MIT-Plato Program

Big Data: Technologies and Business Opportunities

3 day program – July 22-24, 2013 – Room E51-372

Prof. Stuart Madnick

DAY ONE – Introductions and Overviews

1. Introduction to Big Data and Data Quality
2. Mashups & Aggregators: Strategy and Legal/Policy Issues
3. Example of Successful Big Data Company Based on Aggregation: TripAdvisor

- *First Assignment* … Propose Improvements to a Business

DAY TWO – New technologies for Big Data

4. Cloud Computing & Big Data Processing
5. Using Big Data via Web Services to Connect the "Edge of the Organization"
6. Emergence of the Semantic / Intelligent Web & Linked Data

- *Second Assignment* … Develop Proposal for New Business Opportunity

DAY THREE - The future of Big Data and its impact on the world

7. Semantic Representation & Semantic Reasoning
8. Creative Big Data Applications & Course Summary

- *Third Assignment*: Student Presentations
  (Two sessions – depending on number of students and team)
  [e.g., if 20 students = 10 teams of two, would be five 20-minute presentations in each session]

Description of Sessions and Reading List

1 Introduction to Big Data and Data Quality

Introduction to the course. Class will discuss the characteristics Big Data and current trends and the relationship to Data Quality.

a) James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, May 2011, pp. 1-20.
[http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation]

2 Mashups & Aggregators: Strategy and Legal / Policy Issues

Mashups have recently become the basis of many new business models and online services.

a) "Mashup (web application hybrid)," Wikipedia, (last modified on 31 May 2013), 8 pages (in print format).
[http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)]

b) Stuart Madnick and Michael Siegel, "Seize the Opportunity: Exploiting Web Aggregation", MISQ Executive, Vol 1, Issue 1, March 2002, pp. 35-46.
[http://web.mit.edu/smadnick/www/wp/2001-13.pdf]

c) Hongwei Zhu and Stuart Madnick, "One Size does not Fit All: Legal Protection for Non-Copyrightable Data", Communications of the ACM, Vol. 52, No. 9, September 2009, pp. 123-128.
[http://web.mit.edu/smadnick/www/wp/2007-04.pdf]

3 Example of Successful Big Data Company Based on Aggregation: TripAdvisor

Big Data provides opportunities to create new businesses and change existing businesses. These can be based on Consumer-consumer; consumer-business, or business-business interactions. This session will discuss a successful example: TripAdvisor.

a) Sramana Mitra, "TripAdvisor: The Web's Strongest Travel Community," ReadWriteWeb, April 25, 2007, 5 pages.
[http://www.readwriteweb.com/archives/tripadvisor_the.php]

b) Nancy Keates, "Deconstructing TripAdvisor," WSJ Weekend Journal, June 1, 2007, 6 pages.
[http://online.wsj.com/article/SB118065569116920710.html]

4 Cloud Computing & Big Data Processing

This session will discuss: what is cloud computing and its origins and technologies and how they can be applied to Big Data.

a) Wikipedia, "Cloud Computing," 9 July 2013, 28 pages.
[http://en.wikipedia.org/wiki/Cloud_computing]

5 Using Big Data via Web Services to Connect the "Edge of the Organization"

This session will discuss: what are web services and the technologies behind them. Some of the key players and their respective platforms will be discussed, as well as perspectives among providers and consumers.

a) John Hagel, John Seely Brown, Dennis Layton-Rodin, "The Secret to Creating Value from Web Services," 3 pages.
[http://www.johnhagel.com/paper_startsimply.pdf]

6 Emergence of the Semantic / Intelligent Web & Linked Data

Discussion of the long-term vision of the “data web” and “semantic web.” Brief explanation of the Semantic Web and its “Layer Cake” of technologies.

a) Tim Berners-Lee, James Hendler, and Ora Lassila, “The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities,” Scientific American, Vol. 284 Issue 5, May 2001, pp. 34-43.
   [http://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf]

b) Cody Burleson, “Introduction to the Semantic Web Vision and Technologies - Part 1 – Overview”, 24 Sept 2007, Semantic Focus blog, 3 pages.
   [http://www.semanticfocus.com/blog/entry/title/introduction-to-the-semantic-web-vision-and-technologies-part-1-overview]

7 Semantic Representation & Semantic Reasoning

An “ontology” is a representation of "what exists" and is based on the meaning (also known as “semantics”) from traditional philosophy. In the Semantic Web, there are ontology languages such as RDF and OWL. Although semantic representation provides meanings, it is semantic reasoning that makes that knowledge useful.

a) David Hay, “Data Modeling, RDF & OWL – Part One: An Introduction to Ontologies,” The Data Administration Newsletter, April 2006, 14 pages.
   [http://www.tdan.com/view-articles/5025]

c) Hongwei Zhu and Stuart Madnick, “Scalable Interoperability Through the Use of COIN Lightweight Ontology,” ODBIS 2005/2006, (M. Collard (Ed.), LNCS 4623, published by Springer-Verlag Berlin Heidelberg, 2007, pp. 37–50.
   [http://web.mit.edu/smadnick/www/wp/2007-13.pdf]

d) Hongwei Zhu and Stuart Madnick, “Improving Data Quality Through Effective Use of Data Semantics”, Data & Knowledge Engineering, Vol. 59, Issue 2, 2006, pp. 460-476.
   [http://ssrn.com/abstract=825650]

8 Creative Big Data Applications & Course Summary

In this session several different Big Data Applications are described.

a) Cambridge Semantics, “Example Semantic Web Applications,” 2013, 6 pages.
   [http://www.cambridgesemantics.com/semantic-university/example-semantic-web-applications]

McKinsey&Company

McKinsey Global Institute

May 2011

Big data: The next frontier for innovation, competition, and productivity

The McKinsey Global Institute

The McKinsey Global Institute (MGI), established in 1990, is McKinsey & Company's business and economics research arm.

MGI's mission is to help leaders in the commercial, public, and social sectors develop a deeper understanding of the evolution of the global economy and to provide a fact base that contributes to decision making on critical management and policy issues.

MGI research combines two disciplines: economics and management. Economists often have limited access to the practical problems facing senior managers, while senior managers often lack the time and incentive to look beyond their own industry to the larger issues of the global economy. By integrating these perspectives, MGI is able to gain insights into the microeconomic underpinnings of the long-term macroeconomic trends affecting business strategy and policy making. For nearly two decades, MGI has utilized this "micro-to-macro" approach in research covering more than 20 countries and 30 industry sectors.

MGI's current research agenda focuses on three broad areas: productivity, competitiveness, and growth; the evolution of global financial markets; and the economic impact of technology. Recent research has examined a program of reform to bolster growth and renewal in Europe and the United States through accelerated productivity growth; Africa's economic potential; debt and deleveraging and the end of cheap capital; the impact of multinational companies on the US economy; technology-enabled business trends; urbanization in India and China; and the competitiveness of sectors and industrial policy.

MGI is led by three McKinsey & Company directors: Richard Dobbs, James Manyika, and Charles Roxburgh. Susan Lund serves as MGI's director of research. MGI project teams are led by a group of senior fellows and include consultants from McKinsey's offices around the world. These teams draw on McKinsey's global network of industry and management experts and partners. In addition, MGI works with leading economists, including Nobel laureates, who act as advisers to MGI projects.

The partners of McKinsey & Company fund MGI's research, which is not commissioned by any business, government, or other institution.

Further information about MGI and copies of MGI's published reports can be found at www.mckinsey.com/mgi.

McKinsey Global Institute

May 2011

Big data: The next frontier for innovation, competition, and productivity

James Manyika
Michael Chui
Brad Brown
Jacques Bughin
Richard Dobbs
Charles Roxburgh
Angela Hung Byers

Big data—a growing torrent

$600 to buy a disk drive that can store all of the world's music

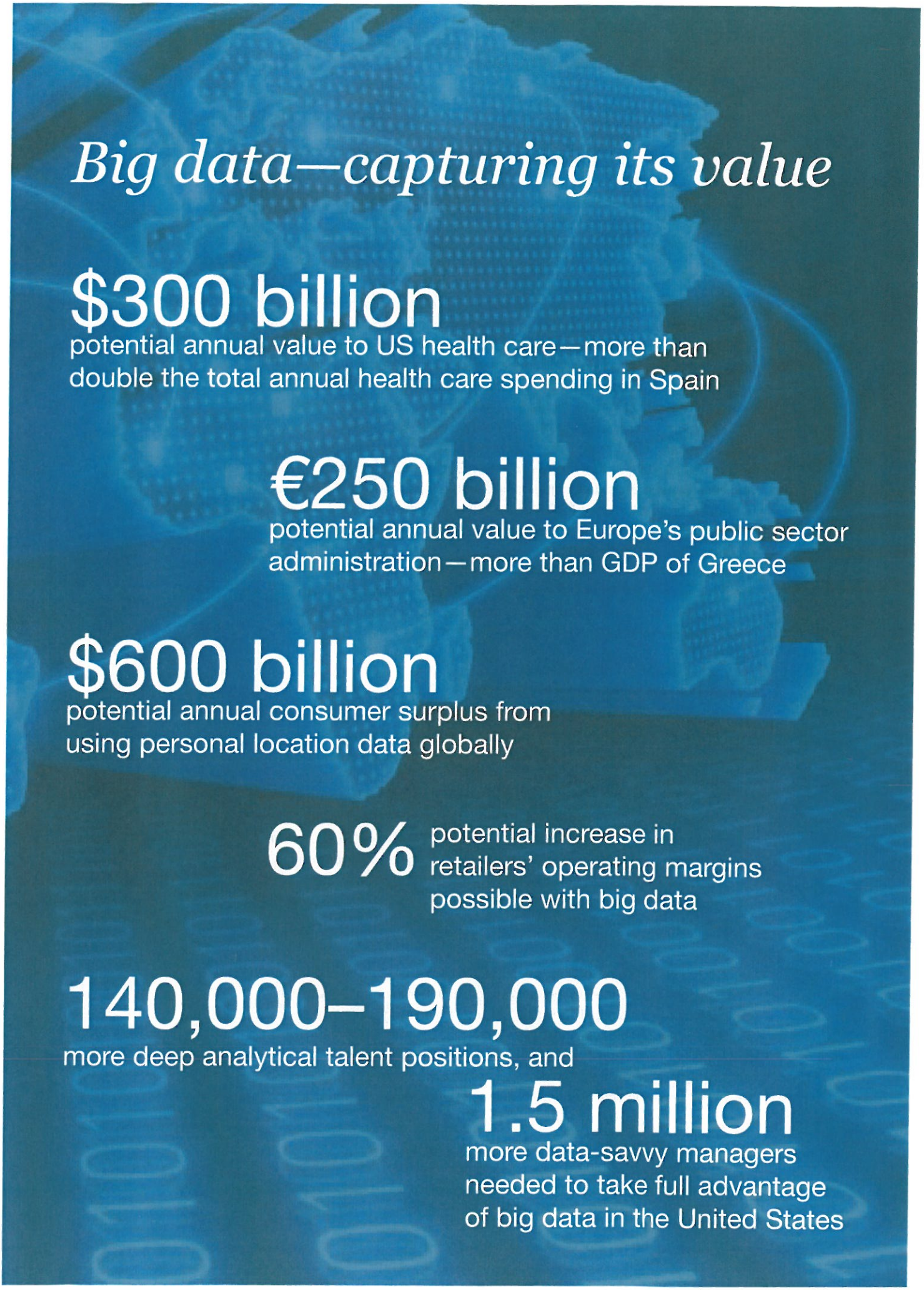
5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. **5%** growth in global IT spending

235 terabytes data collected by the US Library of Congress in April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress


Big data—capturing its value
$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain
€250 billion
potential annual value to Europe's public sector administration—more than GDP of Greece
$600 billion
potential annual consumer surplus from using personal location data globally
60%
potential increase in retailers' operating margins possible with big data
140,000–190,000
more deep analytical talent positions, and
1.5 million
more data-savvy managers needed to take full advantage of big data in the United States

Executive summary

Data have become a torrent flowing into every area of the global economy.[1] Companies churn out a burgeoning volume of transactional data, capturing trillions of bytes of information about their customers, suppliers, and operations. millions of networked sensors are being embedded in the physical world in devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create, and communicate data in the age of the Internet of Things.[2] Indeed, as companies and organizations go about their business and interact with individuals, they are generating a tremendous amount of digital "exhaust data," i.e., data that are created as a by-product of other activities. Social media sites, smartphones, and other consumer devices including PCs and laptops have allowed billions of individuals around the world to contribute to the amount of big data available. And the growing volume of multimedia content has played a major role in the exponential growth in the amount of big data (see Box 1, "What do we mean by 'big data'?"). Each second of high-definition video, for example, generates more than 2,000 times as many bytes as required to store a single page of text. In a digitized world, consumers going about their day—communicating, browsing, buying, sharing, searching—create their own enormous trails of data.

Box 1. What do we mean by "big data"?

"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don't define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).

In itself, the sheer volume of data is a global phenomenon—but what does it mean? Many citizens around the world regard this collection of information with deep suspicion, seeing the data flood as nothing more than an intrusion of their privacy. But there is strong evidence that big data can play a significant economic role to the benefit not only of private commerce but also of national economies and their citizens. Our research finds that data can create significant value for the world economy, enhancing the productivity and competitiveness of companies and the

1 See "A special report on managing information: Data, data everywhere," *The Economist*, February 25, 2010; and special issue on "Dealing with data," *Science*, February 11, 2011.

2 "Internet of Things" refers to sensors and actuators embedded in physical objects, connected by networks to computers. See Michael Chui, Markus Löffler, and Roger Roberts, "The Internet of Things," *McKinsey Quarterly*, March 2010.

public sector and creating substantial economic surplus for consumers. For instance, if US health care could use big data creatively and effectively to drive efficiency and quality, we estimate that the potential value from data in the sector could be more than $300 billion in value every year, two-thirds of which would be in the form of reducing national health care expenditures by about 8 percent. In the private sector, we estimate, for example, that a retailer using big data to the full has the potential to increase its operating margin by more than 60 percent. In the developed economies of Europe, we estimate that government administration could save more than €100 billion ($149 billion) in operational efficiency improvements alone by using big data. This estimate does not include big data levers that could reduce fraud, errors, and tax gaps (i.e., the gap between potential and actual tax revenue).

Digital data is now everywhere—in every sector, in every economy, in every organization and user of digital technology. While this topic might once have concerned only a few data geeks, big data is now relevant for leaders across every sector, and consumers of products and services stand to benefit from its application. The ability to store, aggregate, and combine data and then use the results to perform deep analyses has become ever more accessible as trends such as Moore's Law in computing, its equivalent in digital storage, and cloud computing continue to lower costs and other technology barriers.[3] For less than $600, an individual can purchase a disk drive with the capacity to store all of the world's music.[4] The means to extract insight from data are also markedly improving as software available to apply increasingly sophisticated techniques combines with growing computing horsepower. Further, the ability to generate, communicate, share, and access data has been revolutionized by the increasing number of people, devices, and sensors that are now connected by digital networks. In 2010, more than 4 billion people, or 60 percent of the world's population, were using mobile phones, and about 12 percent of those people had smartphones, whose penetration is growing at more than 20 percent a year. More than 30 million networked sensor nodes are now present in the transportation, automotive, industrial, utilities, and retail sectors. The number of these sensors is increasing at a rate of more than 30 percent a year.

There are many ways that big data can be used to create value across sectors of the global economy. Indeed, our research suggests that we are on the cusp of a tremendous wave of innovation, productivity, and growth, as well as new modes of competition and value capture—all driven by big data as consumers, companies, and economic sectors exploit its potential. But why should this be the case now? Haven't data always been part of the impact of information and communication technology? Yes, but our research suggests that the scale and scope of changes that big data are bringing about are at an inflection point, set to expand greatly, as a series of technology trends accelerate and converge. We are already seeing visible changes in the economic landscape as a result of this convergence.

Many pioneering companies are already using big data to create value, and others need to explore how they can do the same if they are to compete. Governments, too, have a significant opportunity to boost their efficiency and the value for money

---

3 Moore's Law, first described by Intel cofounder Gordon Moore, states that the number of transistors that can be placed on an integrated circuit doubles approximately every two years. In other words, the amount of computing power that can be purchased for the same amount of money doubles about every two years. Cloud computing refers to the ability to access highly scalable computing resources through the Internet, often at lower prices than those required to install on one's own computers because the resources are shared across many users.

4 Kevin Kelly, Web 2.0 Expo and Conference, March 29, 2011. Video available at: www.web2expo.com/webexsf2011/public/schedule/proceedings.

they offer citizens at a time when public finances are constrained—and are likely to remain so due to aging populations in many countries around the world. Our research suggests that the public sector can boost its productivity significantly through the effective use of big data.

However, companies and other organizations and policy makers need to address considerable challenges if they are to capture the full potential of big data. A shortage of the analytical and managerial talent necessary to make the most of big data is a significant and pressing challenge and one that companies and policy makers can begin to address in the near term. The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings. The shortage of talent is just the beginning. Other challenges we explore in this report include the need to ensure that the right infrastructure is in place and that incentives and competition are in place to encourage continued innovation; that the economic benefits to users, organizations, and the economy are properly understood; and that safeguards are in place to address public concerns about big data.

This report seeks to understand the state of digital data, how different domains can use large datasets to create value, the potential value across stakeholders, and the implications for the leaders of private sector companies and public sector organizations, as well as for policy makers. We have supplemented our analysis of big data as a whole with a detailed examination of five domains (health care in the United States, the public sector in Europe, retail in the United States, and manufacturing and personal location data globally). This research by no means represents the final word on big data; instead, we see it as a beginning. We fully anticipate that this is a story that will continue to evolve as technologies and techniques using big data develop and data, their uses, and their economic benefits grow (alongside associated challenges and risks). For now, however, our research yields seven key insights:

1. DATA HAVE SWEPT INTO EVERY INDUSTRY AND BUSINESS FUNCTION AND ARE NOW AN IMPORTANT FACTOR OF PRODUCTION

Several research teams have studied the total amount of data generated, stored, and consumed in the world. Although the scope of their estimates and therefore their results vary, all point to exponential growth in the years ahead.[5] MGI estimates that enterprises globally stored more than 7 exabytes of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks. One exabyte of data is the equivalent of more than 4,000 times the information stored in the US Library of Congress.[6] Indeed, we are generating so much

5 See Peter Lyman and Hal Varian, *How much information? 2003*, School of Information Management and Systems, University of California at Berkeley, 2003; papers from the IDC Digital Universe research project, sponsored by EMC, including *The expanding digital universe*, March 2007; *The diverse and exploding digital universe*, March 2008; *As the economy contracts, the digital universe expands*, May 2009, and *The digital universe decade—Are you ready?*, May 2010 (www.emc.com/leadership/programs/digital-universe.htm); two white papers from the University of California, San Diego, Global Information Industry Center: Roger Bohn and James Short, *How much information? 2009: Report on American consumers*, January 2010, and Roger Bohn, James Short, and Chaitanya Baru, *How much information? 2010: Report on enterprise server information*, January 2011; and Martin Hilbert and Priscila López, "The world's technological capacity to store, communicate, and compute information," *Science*, February 10, 2011.

6 According to the Library of Congress Web site, the US Library of Congress had 235 terabytes of storage in April 2011.

data today that it is physically impossible to store it all.[7] Health care providers, for instance, discard 90 percent of the data that they generate (e.g., almost all real-time video feeds created during surgery).

Big data has now reached every sector in the global economy. Like other essential factors of production such as hard assets and human capital, much of modern economic activity simply couldn't take place without it. We estimate that by 2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data (twice the size of US retailer Wal-Mart's data warehouse in 1999) per company with more than 1,000 employees. Many sectors had more than 1 petabyte in mean stored data per company. In total, European organizations have about 70 percent of the storage capacity of the entire United States at almost 11 exabytes compared with more than 16 exabytes in 2010. Given that European economies are similar to each other in terms of their stage of development and thus their distribution of firms, we believe that the average company in most industries in Europe has enough capacity to store and manipulate big data. In contrast, the per capita data intensity in other regions is much lower. This suggests that, in the near term at least, the most potential to create value through the use of big data will be in the most developed economies. Looking ahead, however, there is huge potential to leverage big data in developing economies as long as the right conditions are in place. Consider, for instance, the fact that Asia is already the leading region for the generation of personal location data simply because so many mobile phones are in use there. More mobile phones—an estimated 800 million devices in 2010—are in use in China than in any other country. Further, some individual companies in developing regions could be far more advanced in their use of big data than averages might suggest. And some organizations will take advantage of the ability to store and process data remotely.

The possibilities of big data continue to evolve rapidly, driven by innovation in the underlying technologies, platforms, and analytic capabilities for handling data, as well as the evolution of behavior among its users as more and more individuals live digital lives.

2. BIG DATA CREATES VALUE IN SEVERAL WAYS

We have identified five broadly applicable ways to leverage big data that offer transformational potential to create value and have implications for how organizations will have to be designed, organized, and managed. For example, in a world in which large-scale experimentation is possible, how will corporate marketing functions and activities have to evolve? How will business processes change, and how will companies value and leverage their assets (particularly data assets)? Could a company's access to, and ability to analyze, data potentially confer more value than a brand? What existing business models are likely to be disrupted? For example, what happens to industries predicated on information asymmetry—e.g., various types of brokers—in a world of radical data transparency? How will incumbents tied to legacy business models and infrastructures compete with agile new attackers that are able to quickly process and take advantage of detailed consumer data that is rapidly becoming available, e.g., what they say in social media or what sensors report they are doing in the world? And what happens when surplus starts shifting from

7 For another comparison of data generation versus storage, see John F. Gantz, David Reinsel, Christopher Chute, Wolfgang Schlichting, John McArthur, Stephen Minton, Irida Xheneti, Anna Toncheva, and Alex Manfrediz, "The expanding digital universe," IDC white paper, sponsored by EMC, March 2007.

suppliers to customers, as they become empowered by their own access to data, e.g., comparisons of prices and quality across competitors?

Creating transparency

Simply making big data more easily accessible to relevant stakeholders in a timely manner can create tremendous value. In the public sector, for example, making relevant data more readily accessible across otherwise separated departments can sharply reduce search and processing time. In manufacturing, integrating data from R&D, engineering, and manufacturing units to enable concurrent engineering can significantly cut time to market and improve quality.

Enabling experimentation to discover needs, expose variability, and improve performance

As they create and store more transactional data in digital form, organizations can collect more accurate and detailed performance data (in real or near real time) on everything from product inventories to personnel sick days. IT enables organizations to instrument processes and then set up controlled experiments. Using data to analyze variability in performance—that which either occurs naturally or is generated by controlled experiments—and to understand its root causes can enable leaders to manage performance to higher levels.

Segmenting populations to customize actions

Big data allows organizations to create highly specific segmentations and to tailor products and services precisely to meet those needs. This approach is well known in marketing and risk management but can be revolutionary elsewhere—for example, in the public sector where an ethos of treating all citizens in the same way is commonplace. Even consumer goods and service companies that have used segmentation for many years are beginning to deploy ever more sophisticated big data techniques such as the real-time microsegmentation of customers to target promotions and advertising.

Replacing/supporting human decision making with automated algorithms

Sophisticated analytics can substantially improve decision making, minimize risks, and unearth valuable insights that would otherwise remain hidden. Such analytics have applications for organizations from tax agencies that can use automated risk engines to flag candidates for further examination to retailers that can use algorithms to optimize decision processes such as the automatic fine-tuning of inventories and pricing in response to real-time in-store and online sales. In some cases, decisions will not necessarily be automated but augmented by analyzing huge, entire datasets using big data techniques and technologies rather than just smaller samples that individuals with spreadsheets can handle and understand. Decision making may never be the same; some organizations are already making better decisions by analyzing entire datasets from customers, employees, or even sensors embedded in products.

Innovating new business models, products, and services

Big data enables companies to create new products and services, enhance existing ones, and invent entirely new business models. Manufacturers are using data obtained from the use of actual products to improve the development of the next generation of products and to create innovative after-sales service offerings. The emergence of real-time location data has created an entirely new set of location-

based services from navigation to pricing property and casualty insurance based on where, and how, people drive their cars.

3. USE OF BIG DATA WILL BECOME A KEY BASIS OF COMPETITION AND GROWTH FOR INDIVIDUAL FIRMS

The use of big data is becoming a key way for leading companies to outperform their peers. For example, we estimate that a retailer embracing big data has the potential to increase its operating margin by more than 60 percent. We have seen leading retailers such as the United Kingdom's Tesco use big data to capture market share from its local competitors, and many other examples abound in industries such as financial services and insurance. Across sectors, we expect to see value accruing to leading users of big data at the expense of laggards, a trend for which the emerging evidence is growing stronger.[8] Forward-thinking leaders can begin to aggressively build their organizations' big data capabilities. This effort will take time, but the impact of developing a superior capacity to take advantage of big data will confer enhanced competitive advantage over the long term and is therefore well worth the investment to create this capability. But the converse is also true. In a big data world, a competitor that fails to sufficiently develop its capabilities will be left behind.

Big data will also help to create new growth opportunities and entirely new categories of companies, such as those that aggregate and analyze industry data. Many of these will be companies that sit in the middle of large information flows where data about products and services, buyers and suppliers, and consumer preferences and intent can be captured and analyzed. Examples are likely to include companies that interface with large numbers of consumers buying a wide range of products and services, companies enabling global supply chains, companies that process millions of transactions, and those that provide platforms for consumer digital experiences. These will be the big-data-advantaged businesses. More businesses will find themselves with some kind of big data advantage than one might at first think. Many companies have access to valuable pools of data generated by their products and services. Networks will even connect physical products, enabling those products to report their own serial numbers, ship dates, number of times used, and so on.

Some of these opportunities will generate new sources of value; others will cause major shifts in value within industries. For example, medical clinical information providers, which aggregate data and perform the analyses necessary to improve health care efficiency, could compete in a market worth more than $10 billion by 2020. Early movers that secure access to the data necessary to create value are likely to reap the most benefit (see Box 2, "How do we measure the value of big data?"). From the standpoint of competitiveness and the potential capture of value, all companies need to take big data seriously. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value. Indeed, we found early examples of such use of data in every sector we examined.

8 Erik Brynjolfsson, Lorin M. Hitt, and Heekyung Hellen Kim, Strength in numbers: *How does data-driven decisionmaking affect firm performance?*, April 22, 2011, available at SSRN (ssrn.com/abstract=1819486).

Box 2. How do we measure the value of big data?

When we set out to size the potential of big data to create value, we considered only those actions that essentially depend on the use of big data—i.e., actions where the use of big data is necessary (but usually not sufficient) to execute a particular lever. We did not include the value of levers that consist only of automation but do not involve big data (e.g., productivity increases from replacing bank tellers with ATMs). Note also that we include the gross value of levers that require the use of big data. We did not attempt to estimate big data's relative contribution to the value generated by a particular lever but rather estimated the total value created.

4. THE USE OF BIG DATA WILL UNDERPIN NEW WAVES OF PRODUCTIVITY GROWTH AND CONSUMER SURPLUS

Across the five domains we studied, we identified many big data levers that will, in our view, underpin substantial productivity growth (Exhibit 1). These opportunities have the potential to improve efficiency and effectiveness, enabling organizations both to do more with less and to produce higher-quality outputs, i.e., increase the value-added content of products and services.[9] For example, we found that companies can leverage data to design products that better match customer needs. Data can even be leveraged to improve products as they are used. An example is a mobile phone that has learned its owner's habits and preferences, that holds applications and data tailored to that particular user's needs, and that will therefore be more valuable than a new device that is not customized to a user's needs.[10] Capturing this potential requires innovation in operations and processes. Examples include augmenting decision making—from clinical practice to tax audits—with algorithms as well as making innovations in products and services, such as accelerating the development of new drugs by using advanced analytics and creating new, proactive after-sales maintenance service for automobiles through the use of networked sensors. Policy makers who understand that accelerating productivity within sectors is the key lever for increasing the standard of living in their economies as a whole need to ease the way for organizations to take advantage of big data levers that enhance productivity.

We also find a general pattern in which customers, consumers, and citizens capture a large amount of the economic surplus that big data enables—they are both direct and indirect beneficiaries of big-data-related innovation.[11] For example, the use of big data can enable improved health outcomes, higher-quality civic engagement with government, lower prices due to price transparency, and a better match between products and consumer needs. We expect this trend toward enhanced consumer surplus to continue and accelerate across all sectors as they deploy big data. Take the area of personal location data as illustration. In this area, the use of real-time traffic information to inform navigation will create a quantifiable consumer surplus through

9 Note that the effectiveness improvement is not captured in some of the productivity calculations because of a lack of precision in some metrics such as improved health outcomes or better matching the needs of consumers with goods in retail services. Thus, in many cases, our productivity estimates are likely to be conservative.

10 Hal Varian has described the ability of products to leverage data to improve with use as "product kaizen." See Hal Varian, *Computer mediated transactions*, 2010 Ely Lecture at the American Economics Association meeting, Atlanta, Georgia.

11 Professor Erik Brynjolfsson of the Massachusetts Institute of Technology has noted that the creation of large amounts of consumer surplus, not captured in traditional economic metrics such as GDP, is a characteristic of the deployment of IT.

savings on the time spent traveling and on fuel consumption. Mobile location-enabled applications will create surplus from consumers, too. In both cases, the surplus these innovations create is likely to far exceed the revenue generated by service providers. For consumers to benefit, policy makers will often need to push the deployment of big data innovations.

Exhibit 1

Big data can generate significant financial value across sectors

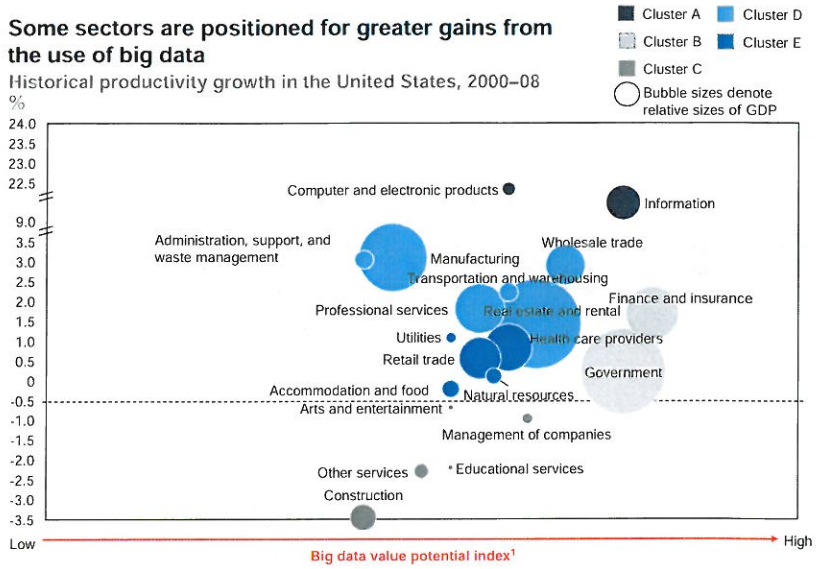
SOURCE: McKinsey Global Institute analysis

5. WHILE THE USE OF BIG DATA WILL MATTER ACROSS SECTORS, SOME SECTORS ARE POISED FOR GREATER GAINS

Illustrating differences among different sectors, if we compare the historical productivity of sectors in the United States with the potential of these sectors to capture value from big data (using an index that combines several quantitative metrics), we observe that patterns vary from sector to sector (Exhibit 2).[12]

12 The index consists of five metrics that are designed as proxies to indicate (1) the amount of data available for use and analysis; (2) variability in performance; (3) number of stakeholders (customers and suppliers) with which an organization deals on average; (4) transaction intensity; and (5) turbulence inherent in a sector. We believe that these are the characteristics that make a sector more or less likely to take advantage of the five transformative big data opportunities. See the appendix for further details.

Exhibit 2



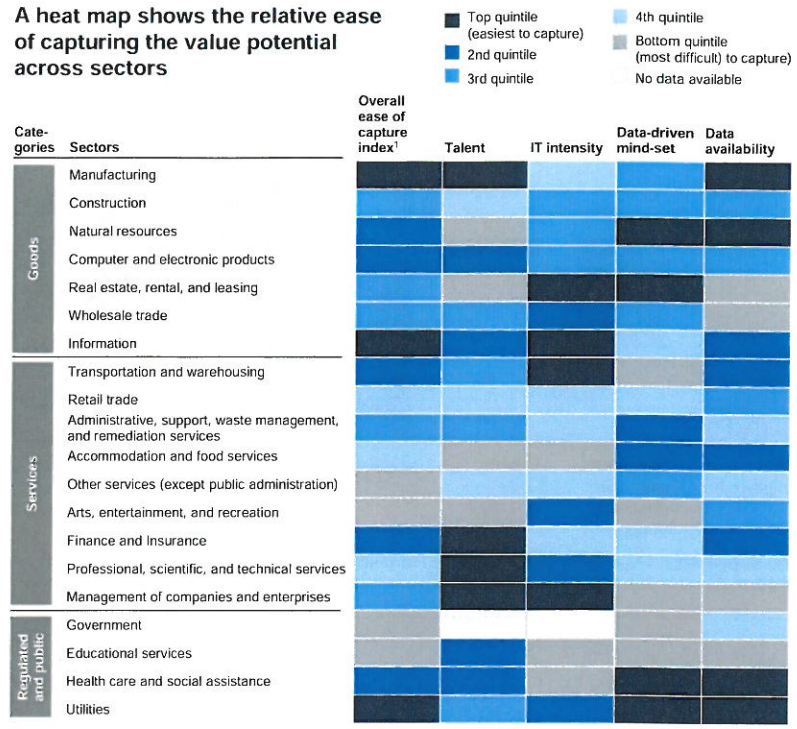
1 See appendix for detailed definitions and metrics used for value potential index.
SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Computer and electronic products and information sectors (Cluster A), traded globally, stand out as sectors that have already been experiencing very strong productivity growth and that are poised to gain substantially from the use of big data. Two services sectors (Cluster B)—finance and insurance and government—are positioned to benefit very strongly from big data as long as barriers to its use can be overcome. Several sectors (Cluster C) have experienced negative productivity growth, probably indicating that these sectors face strong systemic barriers to increasing productivity. Among the remaining sectors, we see that globally traded sectors (mostly Cluster D) tend to have experienced higher historical productivity growth, while local services (mainly Cluster E) have experienced lower growth.

While all sectors will have to overcome barriers to capture value from the use of big data, barriers are structurally higher for some than for others (Exhibit 3). For example, the public sector, including education, faces higher hurdles because of a lack of data-driven mind-set and available data. Capturing value in health care faces challenges given the relatively low IT investment performed so far. Sectors such as retail, manufacturing, and professional services may have relatively lower degrees of barriers to overcome for precisely the opposite reasons.

Exhibit 3

A heat map shows the relative ease of capturing the value potential across sectors



1 See appendix for detailed definitions and metrics used for each of the criteria.
SOURCE: McKinsey Global Institute analysis

6. THERE WILL BE A SHORTAGE OF TALENT NECESSARY FOR ORGANIZATIONS TO TAKE ADVANTAGE OF BIG DATA

A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data.

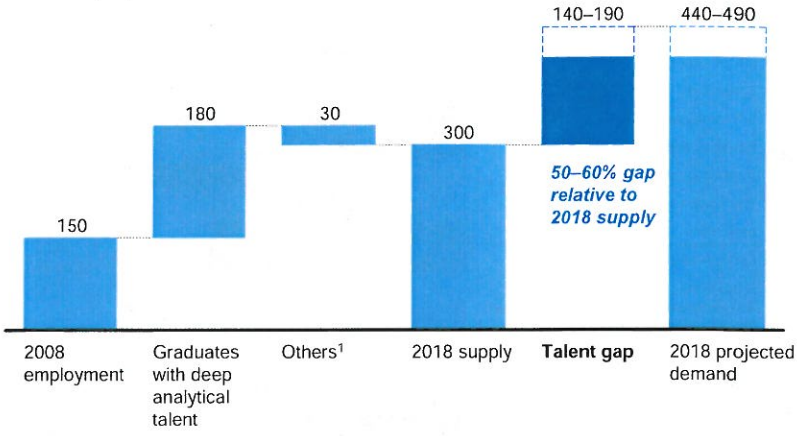
In the United States, we expect big data to rapidly become a key determinant of competition across sectors. But we project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions (Exhibit 4). Furthermore, this type of talent is difficult to produce, taking years of training in the case of someone with intrinsic mathematical abilities. Although our quantitative analysis uses the United States as illustration, we believe that the constraint on this type of talent will be global, with the caveat that some regions may be able to produce the supply that can fill talent gaps in other regions.

In addition, we project a need for 1.5 million additional managers and analysts in the United States who can ask the right questions and consume the results of the analysis of big data effectively. The United States—and other economies facing similar shortages—cannot fill this gap simply by changing graduate requirements and waiting for people to graduate with more skills or by importing talent (although these could be important actions to take). It will be necessary to retrain a significant amount of the talent in place; fortunately, this level of training does not require years of dedicated study.

Exhibit 4

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).
SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

7. SEVERAL ISSUES WILL HAVE TO BE ADDRESSED TO CAPTURE THE FULL POTENTIAL OF BIG DATA

Data policies. As an ever larger amount of data is digitized and travels across organizational boundaries, there is a set of policy issues that will become increasingly important, including, but not limited to, privacy, security, intellectual property, and liability. Clearly, privacy is an issue whose importance, particularly to consumers, is growing as the value of big data becomes more apparent. Personal data such as health and financial records are often those that can offer the most significant human benefits, such as helping to pinpoint the right medical treatment or the most appropriate financial product. However, consumers also view these categories of data as being the most sensitive. It is clear that individuals and the societies in which they live will have to grapple with trade-offs between privacy and utility.

Another closely related concern is data security, e.g., how to protect competitively sensitive data or other data that should be kept private. Recent examples have demonstrated that data breaches can expose not only personal consumer information and confidential corporate information but even national security secrets. With serious breaches on the rise, addressing data security through technological and policy tools will become essential.[13]

Big data's increasing economic importance also raises a number of legal issues, especially when coupled with the fact that data are fundamentally different from many other assets. Data can be copied perfectly and easily combined with other data. The same piece of data can be used simultaneously by more than one person. All of these are unique characteristics of data compared with physical assets. Questions about the intellectual property rights attached to data will have to be answered: Who "owns" a piece of data and what rights come attached with a dataset? What defines "fair use" of data? There are also questions related to liability: Who is responsible when an

13 Data privacy and security are being studied and debated at great length elsewhere, so we have not made these topics the focus of the research reported here.

inaccurate piece of data leads to negative consequences? Such types of legal issues will need clarification, probably over time, to capture the full potential of big data.

Technology and techniques. To capture value from big data, organizations will have to deploy new technologies (e.g., storage, computing, and analytical software) and techniques (i.e., new types of analyses). The range of technology challenges and the priorities set for tackling them will differ depending on the data maturity of the institution. Legacy systems and incompatible standards and formats too often prevent the integration of data and the more sophisticated analytics that create value from big data. New problems and growing computing power will spur the development of new analytical techniques. There is also a need for ongoing innovation in technologies and techniques that will help individuals and organizations to integrate, analyze, visualize, and consume the growing torrent of big data.

Organizational change and talent. Organizational leaders often lack the understanding of the value in big data as well as how to unlock this value. In competitive sectors this may prove to be an Achilles heel for some companies since their established competitors as well as new entrants are likely to leverage big data to compete against them. And, as we have discussed, many organizations do not have the talent in place to derive insights from big data. In addition, many organizations today do not structure workflows and incentives in ways that optimize the use of big data to make better decisions and take more informed action.

Access to data. To enable transformative opportunities, companies will increasingly need to integrate information from multiple data sources. In some cases, organizations will be able to purchase access to the data. In other cases, however, gaining access to third-party data is often not straightforward. The sources of third-party data might not have considered sharing it. Sometimes, economic incentives are not aligned to encourage stakeholders to share data. A stakeholder that holds a certain dataset might consider it to be the source of a key competitive advantage and thus would be reluctant to share it with other stakeholders. Other stakeholders must find ways to offer compelling value propositions to holders of valuable data.

Industry structure. Sectors with a relative lack of competitive intensity and performance transparency, along with industries where profit pools are highly concentrated, are likely to be slow to fully leverage the benefits of big data. For example, in the public sector, there tends to be a lack of competitive pressure that limits efficiency and productivity; as a result, the sector faces more difficult barriers than other sectors in the way of capturing the potential value from using big data. US health care is another example of how the structure of an industry impacts on how easy it will be to extract value from big data. This is a sector that not only has a lack of performance transparency into cost and quality but also an industry structure in which payors will gain (from fewer payouts for unnecessary treatment) from the use of clinical data. However, the gains accruing to payors will be at the expense of the providers (fewer medical activities to charge for) from whom the payors would have to obtain the clinical data. As these examples suggest, organization leaders and policy makers will have to consider how industry structures could evolve in a big data world if they are to determine how to optimize value creation at the level of individual firms, sectors, and economies as a whole.

□ □ □

The effective use of big data has the potential to transform economies, delivering a new wave of productivity growth and consumer surplus. Using big data will become a key basis of competition for existing companies, and will create new competitors who are able to attract employees that have the critical skills for a big data world. Leaders of organizations need to recognize the potential opportunity as well as the strategic threats that big data represent and should assess and then close any gap between their current IT capabilities and their data strategy and what is necessary to capture big data opportunities relevant to their enterprise. They will need to be creative and proactive in determining which pools of data they can combine to create value and how to gain access to those pools, as well as addressing security and privacy issues. On the topic of privacy and security, part of the task could include helping consumers to understand what benefits the use of big data offers, along with the risks. In parallel, companies need to recruit and retain deep analytical talent and retrain their analyst and management ranks to become more data savvy, establishing a culture that values and rewards the use of big data in decision making.

Policy makers need to recognize the potential of harnessing big data to unleash the next wave of growth in their economies. They need to provide the institutional framework to allow companies to easily create value out of data while protecting the privacy of citizens and providing data security. They also have a significant role to play in helping to mitigate the shortage of talent through education and immigration policy and putting in place technology enablers including infrastructure such as communication networks; accelerating research in selected areas including advanced analytics; and creating an intellectual property framework that encourages innovation. Creative solutions to align incentives may also be necessary, including, for instance, requirements to share certain data to promote the public welfare.

Relevant McKinsey Global Institute publications

August 2010
Clouds, big data, and smart assets:
Ten tech-enabled business trends to watch

Advancing technologies and their swift adoption are upending traditional business models. Senior executives need to think strategically about how to prepare their organizations for the challenging new environment.

March 2010
The Internet of Things

More objects are becoming embedded with sensors and gaining the ability to communicate. The resulting new information networks promise to create new business models, improve business processes, and reduce costs and risks.

December 2008
Accounting for the cost of US health care:
A new look at why Americans spend more

The United States spends $650 billion more on health care than expected, even when adjusting for the economy's relative wealth; MGI examines the underlying trends and key drivers of these higher costs. MGI finds that outpatient care, which includes same-day hospital and physician office visits, is by far the largest and fastest-growing part of the US health system. The next largest contributors are the cost of drugs, and administration and insurance.

October 2002
How IT enables productivity growth

Looking at three sectors in detail—retail banking, retail trade, and semiconductors—MGI finds that while IT enabled productivity gains in each sector, its impact was complex and varied. IT applications that had a high impact on productivity shared three characteristics: They were tailored to sector-specific business processes and linked to performance levers; they were deployed in a sequence that allowed companies to leverage their previous IT investments effectively; and they evolved in conjunction with managerial innovation.

October 2001
US productivity growth 1995–2000: Understanding the contributions of information technology relative to other factors

MGI's study of US productivity growth aimed to determine the causes of the sudden increase in the growth rate of labor productivity in the United States between 1995 and 2000. This increase was interpreted by key economists and policy makers as evidence of a "new economy," where IT applications would lead to faster US economic growth. The report lays out the nature and importance of the role played by six key sectors in the acceleration of productivity growth and its causes, with a focus on the role of IT.

www.mckinsey.com/mgi

eBook versions of selected MGI reports are available on MGI's Web site, on Amazon's Kindle bookstore, and on Apple's iBookstore.

Download and listen to MGI podcasts on iTunes or at
www.mckinsey.com/mgi/publications/

McKinsey Global Institute
May 2011

www.mckinsey.com/mgi

Mashup (web application hybrid)

From Wikipedia, the free encyclopedia

A **mashup**, in web development, is a web page, or web application, that uses and combines data, presentation or functionality from two or more sources to create new services. The term implies easy, fast integration, frequently using open application programming interfaces (API) and data sources to produce enriched results that were not necessarily the original reason for producing the raw source data.

The main characteristics of a mashup are combination, visualization, and aggregation. It is important to make existing data more useful, for personal and professional use. To be able to permanently access the data of other services, mashups are generally client applications or hosted online.

In the past years, more and more Web applications have published APIs that enable software developers to easily integrate data and functions instead of building them by themselves. Mashups can be considered to have an active role in the evolution of social software and Web 2.0. Mashup composition tools are usually simple enough to be used by end-users. They generally do not require programming skills and rather support visual wiring of GUI widgets, services and components together. Therefore, these tools contribute to a new vision of the Web, where users are able to contribute.

Contents

History

The history of mashup can be backtracked by first understanding the broader context of the history of the Web. For Web 1.0 business model, companies stored consumer data on portals and updated them regularly. They controlled all the consumer data, and the consumer had to use their products and services to get the information.

With the advent of Web 2.0 a new proposition was created, using Web standards that were commonly and widely adopted across traditional competitors and unlocked the consumer data. At the same time, mashups emerged allowing mixing and matching competitor's API to create new services.

The term isn't formally defined by any standard-setting body.[1]

The first mashups used mapping services or photo services to combine these services with data of any kind and therefore create visualizations of the data.[2] In the beginning, most mashups were consumer-based, but recently the mashup is to be seen as an interesting concept useful also to enterprises. Business mashups can combine existing internal data with external services to create new views on the data.

Mashups are in the ascendant. As a statistic from Programmable Web found out in 2009 that three new mashups have been registered every single day for the last two years.[3]

Types of mashup

There are many types of mashup, such as business mashups, consumer mashups, and data mashups.[4] The most common type of mashup is the consumer mashup, aimed at the general public.

- *Business* (or *enterprise*) *mashups* define applications that combine their own resources, application and data, with other external Web services.[2] They focus data into a single presentation and allow for collaborative action among businesses and developers. This works well for an agile development project, which requires collaboration between the developers and customer (or customer proxy, typically a product manager) for defining and implementing the business requirements. Enterprise mashups are secure, visually rich Web applications that expose actionable information from diverse internal and external information sources.
- *Consumer mashups* combines data from multiple public sources in the browser and organizes it through a simple browser user interface.[5] *(e.g.: Wikipediavision combines Google Map and a Wikipedia API)*
- *Data mashups*, opposite to the consumer mashups, combine similar types of media and information from multiple sources into a single representation. The combination of all these resources create a new and distinct Web service that was not originally provided by either source.

By API type

Mashups can also be categorized by the basic API type they use but any of these can be combined with each other or embedded into other applications.

Data types

- Indexed data (documents, weblogs, images, videos, shopping articles, jobs ...) used by metasearch engines
- Cartographic and geographic data: geolocation software, geovisualization
- Feeds, podcasts: news aggregators

Functions

- Data converters: language translators, speech processing, URL shorteners...
- Communication: email, instant messaging, notification...
- Visual data rendering: information visualization, diagrams
- Security related: electronic payment systems, ID identification...
- Editors

Mashup enabler

In technology, a **mashup enabler** is a tool for transforming incompatible IT resources into a form that allows them to be easily combined, in order to create a mashup. Mashup enablers allow powerful techniques and tools (such as mashup platforms) for combining data and services to be applied to new kinds of resources. An example of a mashup enabler is a tool for creating an RSS feed from a spreadsheet (which cannot easily be used to create a mashup). Many mashup editors include mashup enablers, for example, Presto Mashup Connectors (http://www.jackbe.com/products/connectors.php), Convertigo Web Integrator (http://www.convertigo.com/en/overview/features/web.html) or Caspio Bridge.

Mashup enablers have also been described as "the service and tool providers, that make mashups possible".

History

Early mashups were developed manually by enthusiastic programmers. However, as mashups became more popular, companies began creating platforms for building mashups, which allow designers to visually construct mashups by connecting together mashup components.

Mashup editors have greatly simplified the creation of mashups, significantly increasing the productivity of mashup developers and even opening mashup development to end-users and non-IT experts. Standard components and connectors enable designers to combine mashup resources in all sorts of complex ways with ease. Mashup platforms, however, have done little to broaden the scope of resources accessible by mashups and have not freed mashups from their reliance on well-structured data and open libraries (RSS feeds and public APIs).

Mashup enablers evolved to address this problem, providing the ability to convert other kinds of data and services into mashable resources.

Web resources

Of course, not all valuable data is located within organizations. In fact, the most valuable information for business intelligence and decision support is often external to the organisation. With the emergence of rich internet applications and online Web portals, a wide range of business-critical processes (such as ordering) are becoming available online. Unfortunately, very few of these data sources syndicate content in RSS format and very few of these services provide publicly accessible APIs. Mashup editors therefore solve this problem by providing enablers or connectors.

Data integration challenges

There are a number of challenges to address when integrating data from different sources. The challenges can be classified into four groups: text/data mismatch, object identifiers and schema mismatch, abstraction level mismatch, data accuracy.[6]

Text–data mismatch

A large portion of data is described in text. Human language is often ambiguous - the same company might be referred to in several variations (e.g. IBM, International Business Machines, and Big Blue). The ambiguity makes cross-linking with structured data difficult. In addition, data expressed in human language is difficult to process via software programs. One of the functions of a data integration system is to overcome the mismatch between documents and data.[6]

Object identity and separate schemata

Structured data are available in a plethora of formats. Lifting the data to a common data format is thus the first step. But even if all data is available in a common format, in practice sources differ in how they state what is essentially the same fact. The differences exist both on the level of individual objects and the schema level. As an example for a mismatch on the object level, consider the following: the SEC uses a so-called Central Index Key (CIK) to identify people (CEOs, CFOs), companies, and financial instruments while other sources, such as DBpedia (a structured data version of Wikipedia), use URIs to identify entities. In addition, each source typically uses its own schema and idiosyncrasies for stating what is essentially the same fact. Thus, Methods have to be in place for reconciling different representations of objects and schemata.

Abstraction levels

Data sources provide data at incompatible levels of abstraction or classify their data according to taxonomies pertinent to a certain sector. Since data is being published at different levels of abstraction (e.g. person, company, country, or sector), data aggregated for the individual viewpoint may not match data e.g. from statistical offices. Also, there are differences in geographic aggregation (e.g. region data from one source and country-level data from another). A related issue is the use of local currencies (USD vs. EUR) which have to be reconciled in order to make data from disparate sources comparable and amenable for analysis.

Data quality

Data quality is a general challenge when automatically integrating data from autonomous sources. In an open environment the data aggregator has little to no influence on the data publisher. Data is often erroneous, and combining data often aggravates the problem. Especially when performing reasoning (automatically inferring new data from existing data), erroneous data has potentially devastating impact on the overall quality of the resulting dataset. Hence, a challenge is how data publishers can coordinate in order to fix problems in the data or blacklist sites which do not provide reliable data. Methods and techniques are needed to; check integrity, accuracy, highlight, identify and sanity check, corroborating evidence; assess the probability that a given statement is true, equate weight differences between market sectors or companies; act as clearing houses for raising and settling disputes between competing (and possibly conflicting) data providers and interact with messy erroneous Web data of potentially dubious provenance and quality. In summary, errors in signage, amounts, labeling, and classification can seriously impede the utility of systems operating over such data.

Mashups versus portals

Mashups and portals are both content aggregation technologies. Portals are an older technology designed as an extension to traditional dynamic Web applications, in which the process of converting data content into marked-up Web pages is split into two phases: generation of markup "fragments" and aggregation of the fragments into pages. Each markup fragment is generated by a "portlet", and the portal combines them into a single Web page. Portlets may be hosted locally on the portal server or remotely on a separate server.

Portal technology defines a complete event model covering reads and updates. A request for an aggregate page on a portal is translated into individual read operations on all the portlets that form the page ("`render`" operations on local, JSR 168 portlets or "`getMarkup`" operations on remote, WSRP portlets). If a submit button is pressed on any portlet on a portal page, it is translated into an update operation on that portlet alone (`processAction` on a local portlet or `performBlockingInteraction` on a remote, WSRP portlet). The update is then immediately followed by a read on *all* portlets on the page.

Portal technology is about server-side, presentation-tier aggregation. It cannot be used to drive more robust forms of application integration such as two-phase commit.

Mashups differ from portals in the following respects:

	Portal	Mashup
Classification	Older technology, extension to traditional Web server model using well-defined approach	Using newer, loosely defined "Web 2.0" techniques
Philosophy/approach	Approaches aggregation by splitting role of Web server into two phases: markup generation and aggregation of markup fragments	Uses APIs provided by different content sites to aggregate and reuse the content in another way
Content dependencies	Aggregates presentation-oriented markup fragments (HTML, WML, VoiceXML, etc.)	Can operate on pure XML content and also on presentation-oriented content (e.g., HTML)
Location dependencies	Traditionally, content aggregation takes place on the server	Content aggregation can take place either on the server or on the client
Aggregation style	"Salad bar" style: Aggregated content is presented 'side-by-side' without overlaps	"Melting Pot" style - Individual content may be combined in any manner, resulting in arbitrarily structured hybrid content
Event model	Read and update event models are defined through a specific portlet API	CRUD operations are based on REST architectural principles, but no formal API exists
Relevant standards	Portlet behavior is governed by standards JSR 168, JSR 286 and WSRP, although portal page layout and portal functionality are undefined and vendor-specific	Base standards are XML interchanged as REST or Web Services. RSS and Atom are commonly used. More specific mashup standards such as EMML are emerging.

The portal model has been around longer and has had greater investment and product research. Portal technology is therefore more standardized and mature. Over time, increasing maturity and standardization of mashup technology will likely make it more popular than portal technology because it is more closely associated with Web 2.0 and lately Service-oriented Architectures (SOA).[7] New versions of portal products are expected to eventually add mashup support while still supporting legacy portlet applications. Mashup technologies, in contrast, are not expected to provide support for portal standards.

Business mashups

Mashup uses are expanding in the business environment. Business mashups are useful for integrating business and data services, as business mashups technologies provide the ability to develop new integrated services quickly, to combine internal services with external or personalized information, and to make these services tangible to the business user through user-friendly Web browser interfaces.[8]

Business mashups differ from consumer mashups in the level of integration with business computing environments, security and access control features, governance, and the sophistication of the programming tools (mashup editors) used. Another difference between business mashups and consumer mashups is a growing trend of using business mashups in commercial software as a service (SaaS) offering.

Many of the providers of business mashups technologies have added SOA features.

Architectural aspects of mashups

The architecture of a mashup is divided into three layers:

- Presentation / user interaction: this is the user interface of mashups. The technologies used are HTML/XHTML, CSS, Javascript, Asynchronous Javascript and Xml (Ajax).
- Web Services: the products functionality can be accessed using the API services. The technologies used are XMLHTTPRequest, XML-RPC, JSON-RPC, SOAP, REST.
- Data: handling the data like sending, storing and receiving. The technologies used are XML, JSON, KML.

Architecturally, there are two styles of mashups: Web-based and server-based. Whereas Web-based mashups typically use the user's Web browser to combine and reformat the data, server-based mashups analyze and reformat the data on a remote server and transmit the data to the user's browser in its final form.[9]

Mashups appear to be a variation of a façade pattern.[10] That is: a software engineering design pattern that provides a simplified interface to a larger body of code (in this case the code to aggregate the different feeds with different APIs).

Mashups can be used with software provided as a service (SaaS).

After several years of standards development, mainstream businesses are starting to adopt service-oriented architectures (SOA) to integrate disparate data by making them available as discrete Web services. Web services provide open, standardized protocols to provide a unified means of accessing information from a diverse set of platforms (operating systems, programming languages, applications). These Web services can be reused to provide completely new services and applications within and across organizations, providing business flexibility.

See also

- Open Mashup Alliance
- Web scraping

Notes

1. ^ "Enterprise Mashups: The New Face of Your SOA" (http://soa.sys-con.com/node/719917). http://soa.sys-con.com/: SOA WORLD MAGAZINE. Retrieved 2010-03-03. "The term mashup isn't subject to formal definition by any standards-setting body."
2. ^ ***a*** ***b*** Holmes, Josh. "Enterprise Mashups" (http://msdn.microsoft.com/en-us/architecture/bb906060.aspx). *MSDN Architecture Journal*. MSDN Architecture Center.
3. ^ "Enterprise Mashups: The New Face of Your SOA" (http://soa.sys-con.com/node/719917). http://soa.sys-con.com/: SOA WORLD MAGAZINE. Retrieved 2010-03-03. "One popular mashup site, *Programmable Web*, reports that three new mashups have been registered every single day for the last two years."
4. ^ Sunilkumar Peenikal (2009). "Mashups and the enterprise" (http://www.mphasis.com/pdfs/Mashups_and_the_Enterprise.pdf). MphasiS - HP.
5. ^ "Enterprise Mashups: The New Face of Your SOA" (http://soa.sys-con.com/node/719917). http://soa.sys-con.com/: SOA WORLD MAGAZINE. Retrieved 2010-03-03. "A consumer mashup is an application that combines data from multiple public sources in the browser and organizes it through a simple browser user interface."
6. ^ ***a*** ***b*** E. Curry, A. Harth, and S. O'Riain, "Challenges Ahead for Converging Financial Data," (http://sw.deri.org/2009/09/financial-data/) in Proceedings of the XBRL/W3C Workshop on Improving Access to Financial Data on the Web, 2009.
7. ^ Digna, Larry (2007). "Gartner: The future of portals is mashups, SOA, more aggregation" (http://blogs.zdnet.com/BTL/?p=4912). ZDNET.
8. ^ Holt, Adams (2009). "Executive IT Architect, Mashup business scenarios and patterns" (http://www.ibm.com/developerworks/lotus/library/mashups-patterns-pt1/). IBM DeveloperWorks.
9. ^ Bolim, Michael (2005). "End-User Programming for the Web, MIT MS thesis, 2.91 MB PDF" (http://bolinfest.com/Michael_Bolin_Thesis_Chickenfoot.pdf). pp. 22–23.
10. ^ Design Patterns: Elements of Resuable Object-Oriented Software (ISBN 0-201-63361-2) by Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides

References

- Ahmet Soylu, Felix Mödritscher, Fridolin Wild, Patrick De Causmaecker, Piet Desmet. 2012 . "Mashups by Orchestration and Widget-based Personal Environments: Key Challenges, Solution Strategies, and an Application." (http://www.ahmetsoylu.com/wp-content/uploads/2011/05/Program_Soylu_et_al_2012.pdf) Program: Electronic Library and Information Systems 46 (4): 383–428.

External links

- Why Mashups = (REST + 'Traditional SOA') * Web 2.0 (http://blog.sherifmansour.com/?p=187)
- Mashups Part I: Bringing SOA to the People (http://www.soamag.com/I18/0508-1.asp)
- Mashups Part II: Why SOA Architects Should Care (http://www.soamag.com/I21/0808-1.asp)
- A Mashup with Google Maps and Youtube (http://my-bilingual.com/maps)

MIS Quarterly Executive

SEIZING THE OPPORTUNITY: EXPLOITING WEB AGGREGATION[1]

Stuart Madnick
Massachusetts Institute of Technology

Michael Siegel
Massachusetts Institute of Technology

Executive Summary

There is a new phenomena emerging that can provide significant value to businesses that seize the opportunity or threats for those caught unaware. Web aggregators are entities that collect information from a wide range of web sources, with or without prior arrangements, and add value by providing post-aggregation services. Aggregators have become easier to construct due to new technologies, so we predict they will soon emerge in industries where they do not currently exist – and hundreds already do exist in areas such as financial services, retail, and telecommunications. Like it or not, aggregators will use your web-based information to create new information collections that will affect your current business model, branding, and relationships. Aggregators will change the way your organization operates and the way global e-commerce develops.[2] It is a wise organization that considers its e-strategy, prepares for aggregators, adds aggregation capabilities to its internal and external operations, and fully understands whether it should aggregate or be aggregated.

Why Be Concerned With Web Aggregators?

Imagine you are head of a large, well-established industry giant. Your attitude toward the Internet has shifted from thinking of it as a fad to treating it as an important force in your industry. You have decided to make your product information and ordering available online. After all, your customers are requesting this, and you want to leverage your brand name and your brick-and-mortar assets. After investing heavily in building your online presence, you believe you are ready for this marketplace.

But are you really ready?

On the horizon, unbeknownst to you, is a fast-emerging new entity; it plans to overturn your familiar business landscape. This shopbot-like web aggregator can selectively extract information from your web site, couple it with data from other sources (including your competitors), and handle the necessary fine-tuning to make intelligent comparisons between your and your competitors' offerings.

DealTime.com (see Figure 1) is one example. On a recent comparison-shopping trip, DealTime.com determined that it was less expensive and faster to purchase Reilly and Brown's finance text book, *Investment Analysis and Portfolio Management,* from Amazon.com rather than from A1books.com. If A1books.com's revenue model is based on distributing its products online, the aggregator is likely to dramatically reduce A1books.com's volume and narrow its margins. Furthermore, if Albooks.com's business model is based also on making profits from advertising sales, lead generation fees, or better customer data, the aggregator may be seriously reducing these revenue sources as well.

Aggregators can collect information from *cooperating* and *non-cooperating* sources because new web-based extraction tools allow them to easily and

[1] This article was reviewed and accepted by all the senior editors, including the editor-in-chief. Articles published in future issues will be accepted by just a single senior editor, based on reviews by members of the Editorial Board.

[2] Although many of the cases studied here look mostly at the Business-to-Consumer sector, aggregation activities will play an even more important role in the Business-to-Business side of eBusiness.

Figure 1: Online Book Comparison (Source: www.DealTime.com)

Where Will You Buy?

Investment Analysis and Portfolio Management, 5th edition, by Reilly and Brown, 1996. Hardcover. ISBN: 0030186838. List Price: $107.50.

- Available at A1Books.com for $103.90, including shipping and sales tax, in 5-10 days.
- Available through Amazon.com for $96.79, including shipping and sales tax, in 3-7 days.

transparently gather information from multiple sources with or without the permission or knowledge of the underlying data sources.[3]

Furthermore, aggregators can more easily extract, compare, and analyze information due to the emerging eXtended Markup Language (XML) family of standards (e.g., XML, RDF, XML-Schema). They can also automatically compare information (such as book prices, bank balances, shipping rates, and intelligence information) using mediation technologies, which let them determine differences in semantics or the "meaning" of data.[4] And they can make strategic use of aggregated information using agent technologies, which are programs that use an aggregated information database to perform services on a user's behalf.

Aggregators, by themselves, are not new. What has changed, with the advent of the Internet and recent developments in technology, is their ability to emerge overnight, at minimal cost, and without the need to establish partnerships with the various data sources. As a result, incumbents are often caught off-guard and stumble in their panicked response.

A number of types of aggregators already exist in several industries. They include information management services (to help users manage relationships more effectively), consumer education shopbot services (to compare different products) in the book selling and overnight delivery industries. In their study of a similar phenomenon, which they called "navigators," Evans and Wurster concluded that this is "the battlefield on which competitive advantage will be won or lost."[5] We agree.

What are Web Aggregators?

Here are definitions of a few terms used in this article.

Aggregator

A web aggregator is an entity that can *transparently* collect and *analyze* information from multiple web data sources. In the process, the aggregator resolves the semantic or contextual differences in the information, such as differences in prices extracted from sites that use different currencies or include or exclude shipping charges.

As this definition suggests, web aggregators have three important characteristics:

Access Transparency – An aggregator appears to be a normal user to a data source – simply accessing the information.

Contextual Transparency – An aggregator resolves contextual differences so it can make effective comparisons.

Analysis – Instead of simply presenting data as is, an aggregator uses post-aggregation analysis to synthesize value-added information.

It is important to note that, under this definition, search engines, such as Google and Lycos, and personalized web portals, such as MyNetscape or MyYahoo, are not aggregators. Similarly, web-based malls, category e-stores, and community-based web sites do not fit this category. Although these web sites amass different information, they provide little contextual transparency or analysis.

[3] Firat, A., Madnick, S., and Siegel, M. "The Caméléon Approach to the Interoperability of Web Sources and Traditional Relational Databases," *Proceedings of the Workshop on Information Technology and Systems,* Brisbane, Australia, December 2000; Malchik, A. "An Aggregator Tool for Extraction and Collection of Data from Web Pages," MIT Master's Thesis, 2000.

[4] Goh, C., Bresson, S., Madnick, S., and Siegel, M. "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," *ACM Transactions on Office Information Systems* (17:3), July 1999, 270-293.

[5] Evans, P. and Wurster, T.S. "Getting Real About Virtual Commerce," *Harvard Business Review* (77:6) November 1999, pp. 85-94.

Table 1. Examples of Aggregator Types and Sources

	Comparison	Relationship
Inter-Organizational	Compare book prices or shipping costs of alternative suppliers	Consolidate all one's frequent flyer or financial accounts
Intra-Organizational	Compare manufacturing costs in multiple plants	Consolidate all information about each customer from the company's separately maintained web sites across functions (accounting, service) and geography (domestic and international).

Aggregator Types and Data Sources

Aggregators are used to build *integrated information collections* for many purposes, such as forming *comparisons* and managing *relationships*. These collections can be built from information sources inside an organization (*intra-organizational*), between organizations (*inter-organizational*), or both.

Comparison type aggregators focus on collecting information about specific goods and services for evaluation. Shopbots, used for purchasing books, music, and electronics, are good examples.

Relationship type aggregators form new information collections based on their relationship with aggregatees. For example, financial account aggregators (Yodlee, VerticalOne, CashEdge) are being adopted by major financial institutions (Chase, Citibank, Merrill Lynch) and non-financial institutions (CNBC, AOL).[6] These organizations give their customers the ability to manage all their financial relationships through a single aggregator.[7] Examples of these *aggregator types* and *sources* are shown in Table 1.

As we discuss below, advanced hybrid aggregators can combine several types and sources in a single application.

Aggregatee

An aggregatee is an organization whose information could be collected by an aggregator. Ultimately, aggregators also become aggregatees, because once they provide their services over the web, another aggregator can aggregate their information. We refer to such an aggregator as a *mega-aggregator*. Likewise, as we will see below, many aggregatees may also become aggregators.

After-Aggregation Analysis

After-aggregation, or post-aggregation, refers to the services and analyses applied to a collection of aggregated data. Currently, most aggregators provide a majority of their value merely by creating and providing access to their aggregated information collections (i.e., consolidated financial accounts, frequent flier accounts, competitor prices). But aggregators can extract even greater value from this wealth of information through after-aggregation analysis. For example, although it is very interesting to view all one's financial accounts in a single online report, the real value of such a collection comes from the ability to provide advice (e.g., asset allocation) or to act on the information as an agent for the account owner (e.g., automatically move money from one account to another to maximize return). Finally, privacy issues aside, the owner of an aggregator (i.e., the entity that offers the aggregation service) has valuable information it can use to selectively offer products, tailor marketing, and better understand its business.

Aggregation Examples

One of the best ways to understand aggregation is through examples. Here are several aggregation ex-

[6] As a measure of the projected impact of aggregation, two studies predict a high penetration of account aggregation services: Marenzi, O. "Account Aggregators, Screen Scrappers and Online Financial Services," Report by Celent Communications, March 2000; "Account Aggregation 2.0," *Online Banking Report*, Issue 63, August 2000.

[7] Pan, H. "Integrating Financial Data over the Internet", MIT Masters Thesis, 1999.

Figure 2: From the US Airways Website (emphasis added)

> "US Airways provides Dividend Miles account information for the benefit of its Dividend Miles members. Access to this information is subject to the rules in the Dividend Miles Membership Guide and the liability limitations provided for this website. In addition, by using this site to access your Dividend Miles account, you agree that you will use this site in a manner consistent with the Dividend Miles Membership Guide and **you further agree not to allow access to this site to any third party by revealing your access code to any third party for any reason**. Failure to comply with the foregoing restrictions on the use of this site shall be grounds, in US Airways' sole discretion, for the termination of your access to this site and/or your membership in the Dividend Miles program."

amples with different capabilities. These examples will be very useful later in presenting aggregation opportunities and strategic options.

Relationship Aggregation: Managing Reward Programs via MaxMiles

MaxMiles (www.maxmiles.com) runs a web-based reward management program to help frequent travelers better manage the rewards they earn from airlines, hotels, and car rental companies. Users provide their account and personal identification numbers for all their reward programs to MaxMiles and authorize it to access and analyze their data. In return, MaxMiles provides its customers with a consolidated statement that shows, among other things, the number of points they earned for each account and the number of points that will expire at each date. Users of the MaxMiles service immediately benefit. They do not have to manually keep track of a plethora of passwords and they can view all account activities through a single consolidated statement.

In addition to the standard account statement, MaxMiles offers additional after-aggregation services. For example, it can identify flight segments that possibly were not properly credited. It will deduce that some flight segments may not have been properly posted if, for example, the account data does not show an inbound segment for each outbound flight. This is something no individual airline could do if the trip involved multiple airlines. In the not-too-distant future, MaxMiles expects to offer more personalized account statements that help users take advantage of special offers for which they are interested and eligible.

MaxMiles currently provides its service both to businesses and individual consumers. While the specific revenue from each business partner is not disclosed, individual consumers can sign up for MaxMiles for $2.95/month. The following web portals, travel agents, and reward programs have partnered with MaxMiles:

- AOL and Excite offer the MaxMiles service through their web portals.
- Advanced Travel Management, Journey Corp, Internet Travel Network, and Microsoft's Expedia, offer MaxMiles online mileage management reports through their travel agent sites.
- Hyatt Hotel provides the MaxMiles service for its Diamond and Platinum members.
- XTRA On-Line and Sabre integrate the MaxMiles technology into their travel reservation products.

Interestingly, because MaxMiles does not have to partner with the reward programs to serve its clientele, a wide range of relationships has developed. Some reward programs, such as the Hyatt Gold Passport Program, actively partner with MaxMiles by outsourcing the task to reduce costs and leverage the company's technology to better serve its customers. On the other hand, US Airways initially took a defensive and hostile attitude. In its click-wrap contract, the airline explicitly prohibits flyers from revealing their password to a third party (see Figure 2). US Airways intended to prevent MaxMiles from encroaching on its business. MaxMiles countered by requiring users to give it Limited Power-of-Attorney, as part of its registration process.[8]

[8] As of this writing, although there have been several controversies, there are no definitive legal decisions with regard to aggregation. Some of the current legal issues are discussed in Zhu, H., Madnick, S., and Siegel, M. "Information Aggregation - a Value-Added E-Service," *Proceedings of the 5th International Conference on Technology, Policy, and Innovation – Theme: Critical Infrastructures*, Delft, The Netherlands, June 26-29, 2001. A de-

There are a number of important issues to consider about this aggregator. First, MaxMiles interposes itself between customers and frequent flier programs, the aggregatees. This is important because it may require aggregatees to change their business model as the aggregator replaces a direct relationship with their customers. Aggregatees may choose to cooperate and provide data or financing for preferential treatment (e.g., listing special offers on MaxMiles). They may cooperate to get access to strategic data. For example, MaxMiles is gathering knowledge about how everyone flies, rents cars, and stays at hotels. This set of information is extremely valuable to aggregatees. Aggregatees may also choose to outsource their frequent traveler programs. Or they may be more combative and try to limit access to the data. Regardless of their response, aggregators can significantly impact aggregatees' business and change their relationship with customers.

Comparison Aggregation: Selecting a Carrier Through Intershipper

DealTime, as briefly described earlier, provides comparisons of products, such as books. As a different example, Intershipper (www.intershipper.net) demonstrates both price and non-price information-comparison aggregation services. Given a package source, destination, and weight, Intershipper compares shipping options from multiple carriers (e.g., Fedex, UPS, DHL).

Intershipper also has two additional services. First, it provides a list of the closest drop-off centers for all the carriers. This feature is useful to individuals who do not want to wait around for a scheduled pickup. Second, Intershipper shows when a package is estimated/guaranteed to arrive, based on sender's and recipient's zip code and the package's weight. In essence, Intershipper acts as an intelligent assistant, helping users select the best carrier, not just by estimated cost but also by other such factors as expected and guaranteed delivery times. Since the information Intershipper collects is available on the carriers' web sites, Intershipper has not needed to form explicit partnerships with the carriers to provide its services. This case will be discussed in more depth below.

tailed analysis of the legal issues is being produced in a subsequent report.

Combined Relationship and Comparison Aggregation: Universal Financial Aggregator

As a research experiment, in 1998 we developed the Universal Financial Aggregator (UFA), a demonstration aggregator that would provide integrated access to all one's financial accounts that are accessible online. Instead of seeing only individual accounts or only the accounts from a single institution, users could instantaneously view all their financial accounts across multiple institutions through an integrated, personalized balance sheet. In addition, a UFA would help users manage their plethora of logins and passwords. In this regard the UFA was a *relationship* aggregator, similar to MaxMiles.

To illustrate how rapidly aggregation services can emerge, commercial versions of this aggregator, now called Financial Account Aggregators, appeared in late 1999 from companies, such as Yodlee (www.Yodlee.com), VerticalOne (now merged with Yodlee), and CashEdge (www.cashedge.com).[9] In June 2000, Chase, which had been an aggregatee, announced that it would become an aggregator by working with Yodlee, and would provide financial account aggregation services to its customers. Today, such financial account aggregation services are offered by most major financial services institutions (including Citibank, Chase, Wells Fargo, Merrill Lynch, Fleet Bank, and Fidelity) as well as by non-financial institutions (such as Yahoo, AOL).

With a total picture of a user's financial situation, a financial account aggregator can use its knowledge of other financial products to help the user optimize his or her finances. For example, our experimental UFA incorporated a money market comparison aggregator that scoured the Internet for the best interest rates, consistent with the user's aggregated financial status. In fact, since our aggregator also aggregated other money market rate aggregators (i.e., Bankrate.com and Bankquote.com), we called it a *mega-aggregator*. This capability also made the UFA a *comparison* aggregator, similar to DealTime. Its *after-aggregation* service incorporated *analysis* – evaluating potential additional earned interest by moving funds – and it could act as your *agent*, facilitating the movement of funds. So the UFA has been an example that combines *relationship* and *compari-*

[9] Marenzi, O. "Account Aggregators, Screen Scrappers and Online Financial Services," Report by Celent Communications, March 2000; "Account Aggregation 2.0," *Online Banking Report*, Issue 63, August 2000.

son type aggregations. Some of the high-end commercial financial account aggregators have announced their intention to offer such after-aggregation analysis capabilities in the near future.

Aggregators of all types will affect companies in a wide range of industries. We have examined several hundred examples in the retail, telecommunications, and financial services industries.[10] Early aggregators focused on price comparison. Emerging aggregators focus on relationships and creating and analyzing information collections. In addition, from our UFA experiments, we see that much more functionality and value can be provided by combining aggregation types. In many instances, the result will be a relationship aggregator providing added after-aggregation value through comparison aggregators.

Using Aggregation to Improve Business

Today, barriers of entry to new aggregators are much lower because new web-page extraction tools, context-sensitive mediators, and agent technologies have greatly reduced the time, cost, and effort to construct aggregators.[11] Furthermore, organizations do not need aggregation capabilities in-house. Aggregation service providers license or rent the technologies, so non-technology companies can easily incorporate such services. With the advent of the Internet, many firms have outsourced their technology needs to service providers to benefit from the providers' economies of scale. Hyatt Hotels and various travel agents, for example, have licensed the MaxMiles technology instead of building and maintaining their own aggregation services.

Once one company in an industry provides a useful aggregation service, the others are often compelled to follow. For example, when Chase provided free financial account aggregation, most of the other major financial institutions did the same – mostly by licensing or renting the service from such providers as Yodlee and VerticalOne.

There are many ways a business can exploit aggregation opportunities to its benefit. Aggregation can be used to keep customers, acquire new ones, improve information processing efficiency, generate sales leads, leverage existing customer trust, find suppliers, and understand a market.

To keep customers and acquire new ones. To date, one of the major impacts of aggregators has been their ability to add value to customers' online experiences. For example, relationship aggregators build and maintain customer relationships. Financial services organizations would much prefer customers to access accounts through their own web site rather than through an aggregator's – which might be provided by a third party or even a competitor. This is why financial account aggregation is becoming the "ATM machine of the 21st century." If you do not offer it, your customers will go elsewhere.

Organizations that can add even more value via after-aggregation services will differentiate themselves and place themselves in the best position to keep their existing customers and acquire new ones. In the examples of MaxMiles, Intershipper, and Financial Account Aggregation, the customer relationship has proven to offer the greatest opportunities and concerns for aggregatees.

To process information more efficiently. For manufacturers of information goods, such as Bank Rate Monitor (www.bankrate.com), there is an interesting twist. Aggregators may represent a more efficient model of production. Instead of building their information goods by establishing costly agreements with each data source, aggregators can add and integrate new data sources rapidly and without agreements. More importantly, they may collect information in more ingenious ways, such as offering a service and observing consumer buying patterns. New aggregators may, in fact, displace original manufacturers of information goods that do not seize the opportunity.

Even businesses that are not manufacturers of information goods can use aggregation to better manage their information. Relationship aggregators, for example, can support Customer Relationship Management (CRM) applications, or financial account aggregators can manage a multiplicity of bank accounts, checking accounts, credit cards, certificate of deposits, and money market accounts for a business.

To generate sales leads. Partnering with a comparison or relationship aggregator can help businesses increase sales. *Lead generators* "aggregate [users] … according to their profiles, preferences, and other criteria, translate this data into specific product and service needs, and then direct [users] to vendors

[10] Readers interested in additional case studies should visit the Home of Aggregator Research Web site at context2.mit.edu/aggregation. Included at that site is a list of over a hundred aggregators found in Europe, North American, and Asia.
[11] Firat, et al., 2000 ibid; Goh, et al., 1999 ibid.

whose offerings meet those needs."[12] For example, DealTime.com first identifies possible vendors for a desired book, and then it can direct buyers to the best web site to make the purchase. A financial account aggregator could direct individuals to new and more appropriate investment opportunities.

Not only do lead generators provide businesses with additional customers who are ready to buy, they can more importantly help vendors design better personalized products. As Bakos points out, "Increased selling effectiveness comes from being able to design appropriate products to address the needs of individual consumers, and from being able to identify the moment when a customer's purchasing decision is most likely to occur …"[13]

Sales generators can even provide consumers with structured products tailored to their individual needs by transparently creating and managing a custom bundle of offerings for a particular user. In much the same way that investment banks design products to suit a particular company, we will see aggregation businesses providing tailored, bundled products, such as integrated vacation packages that combine travel, hotel, special events, and equipment rentals. As another example, a transaction coordinator can offer college students bundles of textbooks that match their classes, sourcing the books from various sites and coordinating their delivery, all transparently to the students.

To leverage existing customer trust. While trust has always been important in doing business, it will become even more critical in electronic commerce. Absence of face-to-face contact between buyer and seller, and the ease with which a small (or illegitimate) outfit can appear large (and legitimate), puts small, unrecognized new entrants at a great disadvantage. Historically, retailers have provided face-to-face trust for small producers. It makes sense, therefore, for well-known retailers to build or invest in an aggregator and leverage its brand image to *facilitate transactions* through escrow services, quality guarantees, and extensions of credit. CNET's certification program, for instance, automatically extends CNET's name and legitimacy to small and relatively unknown retailers.

To find suppliers. Buyer-oriented aggregators can serve as *purchasing agents*, searching for the best provider. These buyer agents "help [consumers] get maximum value from their information profiles by using choices they have made in the past to deduce which product or service would best match their current needs, and then finding the vendor that can deliver the preferred product or service at the cheapest price."[14] These agents could even create aggregated products.

As MaxMiles illustrates, aggregators can help users manage multiple relationships. More importantly, they can generate more personalized recommendations than individual organizations, once they have the needed personal information. In these cases, buyers can build and maintain their own aggregators, *subscribe to the service* of an aggregator, or even pay aggregators a *commission on savings.* TPN Register (www.tpnregister.com), a joint venture between GE and Thomas Publishing Company, allows buyers to post design and engineering specifications for bids by suppliers. "The system allows its users, especially from smaller companies, to find low bidders among suppliers that might not consider them via traditional channels."[15]

To understand a market. Aggregators are well positioned to collect detailed and highly valuable market information not available to individual aggregatees. By simultaneously accessing and integrating information from multiple sources, aggregators can understand a market better than its participants. While a company's web site can gather information about its customers, it does little to inform the company about its *non-customers*, that is, those who take their business elsewhere.

For example, Intershipper knows which carrier each user ultimately chooses, and it knows which users use UPS for all packages over one-pound between Boston and New York and Fedex for other shipments. The shippers do not have this information. Consequently, aggregators can sell summarized and aggregated information to individual firms. Of course, such information providers existed before the Internet. IMS America collects, aggregates, and repackages data from hospitals for sale back to those same hospitals, so they can see how their operations compare with their peers. As the cost of collecting and integrating information falls, aggregators will increasingly provide after-aggregation market knowledge in different industries.

[12] Hagel III, J. and Rayport, J.F. "The New Infomediaries", *The McKinsey Quarterly* (4), 1997, pp. 54-70.

[13] Y. Bakos, "The Emerging Role of Electronic Marketplaces on the Internet", *Communications of the ACM* (41:8), 1998, pp. 35-42.

[14] Hagel and Rayport, 1997, ibid.

[15] Segev, A., Gebauer, J. and Farber, F. "Internet-based Electronic Markets," *EM - International Journal of Electronic Markets* (9:3), 1999.

Strategic Relationships Between Aggregators and Aggregatees

Based on our observations, aggregators' strategies are often emergent, rather than planned. They can appear as new entrants in an industry or as new divisions in an existing organization. In the initial phase, aggregatees may be just beginning to formulate their online strategy so they are turning themselves into aggregation targets without realizing the consequences of their actions.

Aggregators often emerge quickly and catch aggregatees off-guard. For example, an existing office supply product provider might build an aggregator to obtain market intelligence on competitors' product pricing – without the aggregatees' knowledge.

Then, once the aggregator realizes it might be able to sell that information, it develops a more mature strategy and strengthens its relationship with the aggregatees. Formal partnerships can reduce an aggregator's integration costs, and aggregatees may gladly pay for preferential treatment. In such cases, the aggregator is a "financially independent aggregator with collaboration," while the aggregatees are "collaborating aggregatees."

Aggregatees who view an aggregator's strategy as a threat may develop their own aggregator. Others may seek to control the existing aggregator through ownership. Still others may work with incumbent aggregatees to create a better balance-of-power, if they face a well-funded competitor. In all these cases, the aggregators are financially dependent, either on a single aggregatee or a consortium of aggregatees.

In general, the different states of aggregation can be characterized by (1) the preference given an aggregatee, (2) the amount of financial control over the aggregator, and (3) the number of participants in an agreement. Table 2 summarizes the different relationships. Each is discussed in the Appendix.

Table 2 presents the progression of aggregator/aggregatee relationships in a linear fashion, proceeding from "no aggregator" to independent aggregator to collaborative aggregator. However, these strategic relationships are dynamic and multidimensional. An aggregator can just as easily establish partnerships with or without investment from industry incumbents. Similarly, an aggregator that begins life as a subsidiary of an incumbent can be divested to become a financially independent aggregator.

Table 2: Summary of Different Relationship States Between an Aggregator and an Aggregatee

Aggregator	Aggregatee
No Aggregation	
• Non-aggregator	• Aggregatee – but no aggregation yet
Aggregation Without Partnership	
• Financially Independent Aggregator	• Unsuspecting Aggregatee
Aggregation with Partnership	
• Financially Independent Aggregator with Partial Collaboration	• Collaborating Aggregatee
• Financially Independent Aggregator with Limited Alliance	• Collaborating Aggregatee Member of a Limited Alliance
• Financially Independent Aggregator with Equal Degrees of Collaboration	• Collaborative Aggregatee
Aggregation with Ownership	
• Financially Dependent Aggregator Owned by a Dominant Aggregatee	• Dominant Aggregatee
• Financially Dependent Aggregator Owned by a Consortium of Aggregatee	• Consortium of Aggregatees

Comparing Strategic Interactions: Intershipper Versus iShip

Intershipper provides a good example of how aggregator and aggregatee relationships and business models can evolve over time.

BITS, Inc., the parent of Intershipper, began as an independent company. Its main source of revenues came from selling network equipment online and hosting online storefronts for various merchants. BITS built Intershipper to allow its online storefront customers to rapidly compare shipping prices across multiple shippers, for free. A spread of ten times in shipping rates was not uncommon. Table 3 shows some estimated shipping rates for a one-pound package from Cambridge, Massachusetts to Arlington, Virginia; they vary from $3 to $125. Traditionally, obtaining such comparative rate information was difficult.

Intershipper became an aggregator; the carriers were the aggregatees. When one of the unsuspecting carriers realized what had happened, it became furious and had its corporate counsel write a letter demanding that Intershipper cease and desist from aggregating its information. Since Intershipper had several other carriers it could aggregate and it did not want to incur legal expenses, it agreed to remove the carrier from its list. Some six months later, the carrier's business development managers decided they wanted to be back on Intershipper's listing. So they asked to be readmitted. Intershipper agreed.

BITS realized that Intershipper might be useful to customers beyond its captive online storefronts. To attract users, BITS let them access Intershipper for free, supporting the cost of operations by both selling advertising space and licensing its service for a fee to other web sites that need to ship goods.

Despite the large number of advertising-supported web sites, few earn a profit. Moreover, seeing how the UPS-owned competitor, iShip, was better funded and could possibly compete even at a loss for a much longer period of time, Intershipper needed to change its strategy. This was the situation when we last interviewed Intershipper.

What are Intershipper's options? One is for Intershipper to leverage its position as an intermediary and dole out preferential treatments in return for fees. We believe this is a shortsighted strategy because maintaining biased relationships will encourage other shipping carriers to introduce their own aggregators, which will increase competition.

At the moment, Intershipper contrasts nicely against iShip. Intershipper is an independent aggregator whereas iShip is not. Thus, carriers other than UPS should have a vested interest in supporting Intershipper and its independent status.

As it now stands, UPS has advantages over its competitors because it controls iShip. UPS can determine, for example, the factors, location, and time of comparison, and it knows more about the industry than its competitors. It knows exact conditions – route, price, package, and type of user – under which a particular competitor was selected. We argue this is highly useful market data not available elsewhere.

Intershipper, being an independent aggregator, can provide the same level of information to the other carriers. Instead of each building its own aggregator, we think Intershipper's better option is to get the

Table 3: Some Rates for Sending a One-Pound Package from Cambridge, MA to Arlington, VA (Source: www.intershipper.net)

Carrier Service	Date Delivered	Rate
RPS Ground	8/17 (Guaranteed)	$3.25
UPS Ground (Commercial)	8/17 (Guaranteed)	$3.25
U.S.P.S. Priority Mail with Confirmation	8/16	$3.55
FedEx Priority Overnight w/ Sat. Delivery	8/14 (Guaranteed)	$30.50
UPS Next Day Air Early AM	8/16 by 8:30 AM (Guaranteed)	$43.50
FedEx First Overnight	8/16 by 8:00 AM (Guaranteed)	$45.50
UPS Next Day Air Early AM w/ Sat. Delivery	8/14 by 9:30 AM (Guaranteed)	$53.50
BAX Guaranteed Overnight	8/16 by 5:00 PM (Guaranteed)	$125.00

other carriers to jointly invest in it, to get the benefits UPS enjoys with significantly less risk.

Legal and Policy Issues

Organizations rushing to put their information on the Internet are just beginning to realize the impact of aggregators using that information. Many are not prepared for open comparison with competitors, the disintermediation that can occur, or the lost opportunity from not harvesting competitive information. Senior executives have only recently begun talking about aggregation strategies. Yet, aggregation will play a significant role in most enterprises, both private sector and government.

As a result, legal and political issues are emerging. For example, various types of legislation are under consideration in the U.S. (e.g., Coble Bill, Bliley Bill, Gramm-Leach-Bliley Act), which address who and how web information can be re-used. International laws will also affect the location, operations, and future of aggregators because those not allowed in one country may simply operate in another.

Research is exploring the impact of regional and global legal, economic, and cultural issues on the development of local and global aggregators.[16] The outcome of these domestic and international actions may influence the development of aggregators. But, in spite of some high-profile challenges to some aggregators (e.g., eBay vs. AuctionWatch and Bidder's Edge), most challenges seem doomed to fail simply because customers will demand access to information through aggregators.

Conclusions

Let us go back to "the head of a large and well-established industry giant" introduced at the beginning of this paper. What might his or her organization learn from this discussion? This research demonstrates that *everyone can be impacted* by aggregation. Everyone with useful information on their web sites is likely to *become an aggregatee*. In response, or to preempt the opportunity, some may *become an aggregator* as well. Thus, aggregation strategy must be part of e-business and core business strategic planning.

Aggregation is *not a disappearing dot-com phenomenon*. Aggregators create new and valuable information spaces, important to organizations in many business areas. In fact, in some industries, such as financial services, the key providers of financial aggregation services are the largest, most established companies (e.g., Chase, Citibank, Merrill Lynch).

Although *comparison aggregation* (e.g., DealTime, MySimon) might be the most common type of aggregation service today, other types, especially *relationship aggregation*, are likely to be even more important. Furthermore, as seen with the Universal Financial Aggregator (UFA) example, it is possible to *combine multiple types of aggregators* to provide totally new services.

Because the impact is so widespread and significant, the aggregation phenomena will change, and will continue to change, business relationships and create new partnerships. The need to share information and gain value from these new information spaces will result in both established and newly created organizations working together in new ways. The wealth of knowledge to be garnered from the new information spaces, the after-aggregation analyses, and the new relationships that evolve will change the way organizations do business. Organizations that ignore the potential impact will be hurt by those that take aggregated information into consideration.

Thus companies should look upon aggregation as both a threat and an opportunity. The airline industry should think about what could happen if MaxMiles becomes the primary frequent-flyer aggregator, and thus owns all the information about who flies where and when. Likewise, a computer retailer with no brand recognition should think about becoming a certified merchant of CNET-owned computers.com, to gain a level playing field with retailers that are spending millions of dollars in advertising.

Like it or not, aggregators will use your web-based information to create new information collections that will affect your current business model, branding, and relationships. Aggregators will change the way your organization operates and the way global e-commerce develops. It is a wise organization that considers its e-strategy, prepares for aggregators, adds aggregation capabilities to its internal and ex-

[16] Zhu, H., Madnick, S. and Siegel, M. "Information Aggregation - a Value-Added E-Service", *Proceedings of the 5th International Conference on Technology, Policy, and Innovation – Theme: Critical Infrastructures*, Delft, The Netherlands, June 26-29, 2001.

ternal operations, and fully understands whether it should aggregate or be aggregated.

Acknowledgements

The authors would like to acknowledge the contributions of Steven Chan, Mary Alice Frontini, Saraubh Khemka, and Howard Pan. All these MIT students provided significant contributions to this research and to preliminary versions of this paper. Work reported herein has been supported, in part, by the Advanced Research Projects Agency (ARPA) and the USAF/Rome Laboratory (under contract F30602-93-C-0160), Citibank, Fleet Bank, Merrill Lynch, Suruga Bank, and PricewaterhouseCoopers.

About the Authors

Dr. Stuart E. Madnick (smadnick@mit.edu, http://web.mit.edu/smadnick/www/home.html) is the John Norris Maguire Professor of Information Technology in the Sloan School of Management and a Professor of Engineering Systems in the School of Engineering at the Massachusetts Institute of Technology. Dr. Madnick has served as the Head of the Information Technologies Group for more than twenty years and is the author or co-author of over 250 books, articles, or reports. His current research interests include connectivity among disparate distributed information systems, database technology, software project management, and the strategic use of information technology. He is presently co-Director of the PROductivity From Information Technology (PROFIT) Initiative and the Total Data Quality Management research program. He was a key designer of IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system. He has consulted to many major corporations, has been the founder or co-founder of several high-tech firms, and currently operates a hotel in the 14th century Langley Castle in England. Dr. Madnick has degrees in Electrical Engineering (B.S. and M.S.), Management (M.S.), and Computer Science (Ph.D.) from MIT.

Dr. Michael Siegel (msiegel@mit.edu) is a Principal Research Scientist at the MIT Sloan School of Management. He co-directs the Aggregation Research Project at the Sloan School of Management and is a co-inventor on several related patents including "Querying Heterogeneous Data Sources over a Network Using Context Interchange" and "Data Extraction from World Wide Web Pages." Dr. Siegel's research interests include the use of information technology in financial risk management and global financial systems, eBusiness and financial services, financial account aggregation, heterogeneous database systems, managing data semantics, query optimization, intelligent database systems, and learning in database systems. His work in benchmarking Value-at-Risk software systems has been presented to the Federal Reserve Bank, academic, and international audiences. Dr. Siegel obtained Engineering Degrees from Trinity College and the University of Wisconsin-Madison and a Ph.D. in Computer Science from Boston University.

Appendix: Aggregator / Aggregatee Relationships

No Aggregation

The *No Aggregation* state is the base case and is probably the accustomed state for most firms. Each incumbent with an online presence is an aggregatee, and provides a target for consolidation; but no active aggregators exist yet. The more inefficient the information dissemination and the more difficult it is to compare like products, the more likely an aggregator will emerge to eliminate the inefficiency.

Aggregation Without Partnership

Financially independent aggregator / unsuspecting aggregatee. These aggregators generally access widely available information that can be extracted without an aggregatee's knowledge, so there is no *a priori* need to establish a partnership or arrangement between the two. In fact, aggregatees usually cannot differentiate between normal users accessing their information and an aggregator accessing the information (using a user's password, if necessary).

In Intershipper's case, as noted above, one carrier sent a letter threatening legal action, then changed course. These actions demonstrate that many aggregatees are completely unprepared for aggregation in their industry.

Aggregation with Partnership

Although some aggregatees engage in a hostile relationship with an aggregator, others will choose to

build mutually beneficial partnerships. Such partnerships may facilitate the aggregator's data extraction and also allow it to obtain information not yet on the web. For example, Intershipper has access to publicized shipping rates, but not customer-specific negotiated rates. Partnering with aggregatees is one way for Intershipper to get this data.

In this *aggregation with partnership* space, the entities can have bilateral relationships negotiated one-to-one, or they can opt for an industry-wide relationship with equal treatment to all. Or a selective group can build a limited alliance, with only specific aggregatees as members. Depending on the relative sizes of the aggregatees, the fragmentation of the industry, and antitrust concerns, one form of partnership may be preferable to another.

Financially independent aggregator with partial collaboration / collaborating aggregatee. To differentiate a relationship, an aggregator may leverage its intermediary position and give preferential treatment to an aggregatee in return for a fee. Or an aggregatee may differentiate itself from its competitors through a special relationship. For example, on its Computers.com Web site, CNET differentiates individual retailers through a certification process. CNET-certified retailers receive preferential listings and may appear more credible to consumers.

Financially independent aggregator of a limited alliance / collaborating aggregatee member of a limited alliance. When an industry has a high degree of rivalry, the participants may avoid partnerships with competitors. Aggregatees may seek to sharply limit an aggregator's list of potential partners.

Financially independent aggregator with equal degrees of collaboration / collaborative aggregatee. On the other hand, an aggregator may value its long-run neutrality over short-term gains from doling out preferential treatments. Such aggregators that want to serve as electronic marketplaces or in an advisory role must maintain their impartiality at all times. They are likely to provide equal collaboration to all aggregatees.

Aggregation with Ownership

Similarly, aggregatees may decide to strengthen and lock in their partnership with an aggregator through direct investment. Again, the options parallel those before: an aggregatee can form a consortium to invest in the aggregator or invest on its own.

Financially dependent aggregator owned by a dominant aggregatee / dominant aggregatee. An aggregatee can decide to invest in an existing aggregator or even preemptively launch its own aggregator. For example, UPS decided to launch its own aggregator called iShip. This allows UPS to maintain more control over who is included as its competitor, how UPS will be compared against them, and how the comparison will be made. By owning the aggregator, UPS can access information about how users of the aggregator ship. This can provide UPS with a tremendous strategic advantage.

Financially dependent aggregator owned by a consortium of aggregatees / consortium aggregatee. To counteract the possibility of a single aggregatee dominating an aggregator, a group of aggregatees may form a consortium and make equal investments into an independent aggregator. For example, three large steel manufacturers – LTV Steel, Steel Dynamics, and Weirton Steel – built Metal Site (metalsite.net) as a neutral marketplace for their industry. This action eliminates competitive bidding for the aggregator's preferential treatment and provides the consortium of aggregatees with control over the aggregator.[17]

[17] Segev, et al., 1999, ibid.

One Size does not Fit All:
Legal Protection for Non-Copyrightable Data

Hongwei Zhu
Stuart E. Madnick

Working Paper CISL# 2007-04

July 2007

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

One Size does not Fit All: Legal Protection for Non-Copyrightable Data

Hongwei Zhu
College of Business & Public Administration
Old Dominion University
2147 Constant Hall
Norfolk, VA 23529
USA
hzhu@odu.edu

Stuart E. Madnick
Sloan School of Management
Massachusetts Institute of Technology
30 Wadsworth Street, E53-321
Cambridge, MA 02142
USA
smadnick@mit.edu

Introduction

The Web has become the largest data repository on the planet[1]. An important factor contributing to its success is its openness and ease of use: anyone can contribute data to, and consume data from, the Web. As Tim Berners-Lee, inventor of the Web, said[2], "the exciting thing is serendipitous reuse of data: one person puts data up there for one thing, and another person uses it another way". Such serendipitous data reuse is extremely valuable. Through reuse, new knowledge can be created, innovation and value-added services become possible.

However, there have been efforts to regulate and legally challenge data reuse activities. The European Union (EU) has adopted the Database Directive to restrict unauthorized data extraction and reuse. In the U.S., Congress has considered six bills, all of which failed to pass into law. These legislative activities are summarized in Figure 1; more details are furnished later. The significant uncertainty and the international differences in database legislation have created serious challenges to the "serendipitous reuse of data". The dual purposes of this paper, both related to the theme "one size does not fit all", are to: (1) summarize the range of legislation in current use and proposed, and (2) present an economic model for interpreting or recommending policy choices that depend on factors such as cost of database creation and level of database differentiation.

[1] In the ensuing discussion, we will consider a website owner as a database creator.
[2] An interview by *Technology Review*, October, 2004, p44.

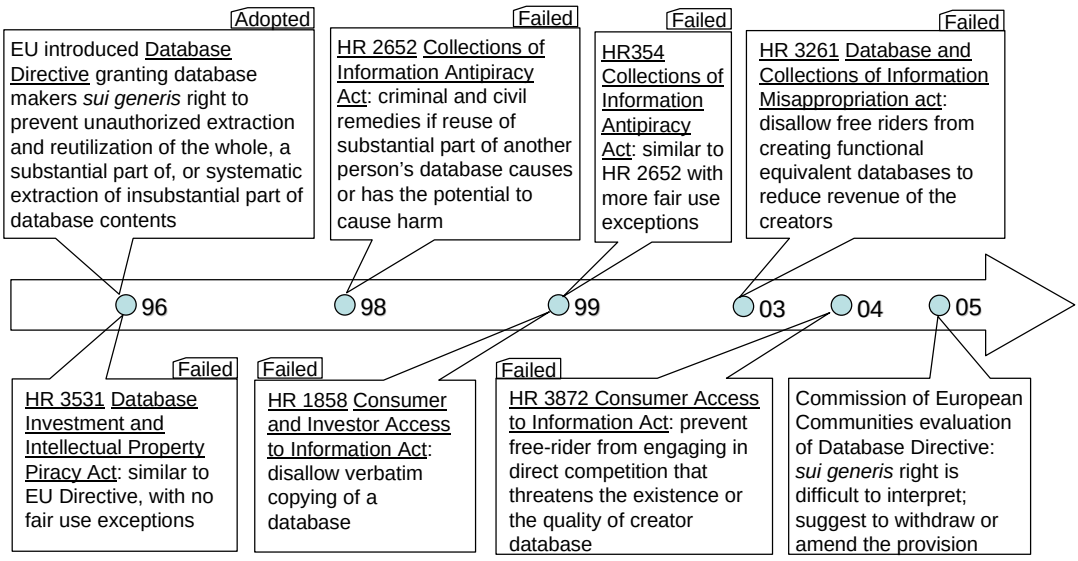


Figure 1. History of Database Protection Legislation

As computing professionals continue to develop technologies (e.g., data extract, web mashups, web services, and various Semantic Web technologies) to make data reuse much easier, it is important for us to understand the legal implications when applying these technologies for data reuse purposes.

eBay *v.* Bidder's Edge: Data Reusers Face Legal Challenges

Let us start with an example. With millions of items auctioned at hundreds of online auction sites, it can be time consuming to find the specific items of interest and keep track of their bidding prices on multiple auction sites. A number of auction data aggregators, such as Bidder's Edge, emerged to address the challenge by employing computer agents to visit auction sites repeatedly and extract data systematically. Bidder's Edge made search and comparison of auction data across multiple sites much easier by gathering bidding data of over five million items from more than 100 online auction sites, including eBay. However, in late 1999, eBay sued Bidder's Edge and won a preliminary injunction in the following year based on a controversial interpretation of

trespass law in the Internet context [9]. The case was settled later without a court decision; Bidder's Edge ceased operation and the company no longer exists.

There have been several other cases involving data reuse in the U.S. A common characteristic in these cases is that the data reusers (e.g., Bidder's Edge) tend to be smaller firms using new technologies to extract and reuse data from one or more creator databases. In many cases, the data reusers stopped their activities in fear of the legal threats posed by the creators. Existing and emerging technology-enabled data reusers continue to face legal challenges. For example, data reusers that provide airfare comparison services have received warning letters from some online travel agencies[3].

Data reusers in Europe have also faced legal challenges. For example, William Hill, an online betting company in the U.K., created a database by combining its own data (e.g., betting odds) with horse racing event data published by British Horseracing Board (BHB), which is the governing authority for organizing horse races in the U.K. William Hill displayed the contents of the database on its website to facilitate its betting business, but was sued by BHB for its systematic reuse of BHB's data.

These cases have raised several questions regarding technology-enabled data reuse: Is it legal? Should it be regulated? If so, what are the issues and how should it be regulated? We will address these questions in the rest of the paper.

Feist *v.* Rural: Non-Creative Database Contents Are Not Copyrightable in the U.S.

Many people think that the factual data on websites is copyrighted, thus extraction and reuse of the data from websites is well-defined and controlled by copyright law. It turns out that is not the case.

[3] See "Cheap-Tickets Sites Try New Tactics" by A. Johnson, Wall Street Journal, October 26, 2004.

When it comes to data, copyright in the U.S.[4] protects the original selection and arrangement of data, but not the data itself or the effort in compiling the database. This principle was established in a landmark Supreme Court case between *Feist Publications* and *Rural Telephone Co*.[5] In compiling its phone book covering the service area of Rural, Feist reused 1,309 of the approximately 7,700 listings in Rural's White Pages. In the appeal case, the Supreme Court decided that Feist did not infringe Rural's copyright in that Rural's white pages lack the requisite originality to warrant copyright protection. Originality requires a work to be "independently created by the author" and it must possess "at least some minimal degree of creativity". Arranging entries alphabetically does not have the required degree of creativity.

The Court confirmed that "copyright rewards originality", originality requires "some minimal degree of creativity", and "Originality is a constitutional requirement." It also rejected the so-called "sweat of the brow" doctrine that considers copyright as a "reward for the hard work that went into compiling facts." The implication of this landmark decision is that in the U.S. copyright currently does not restrict the reuse of the factual contents in most publicly accessible databases on the Web[6].

The Court decision, together with the exponential growth of digital information and the increasing technological capability of reusing information, have induced a series of legislative activities to provide legal protection for database contents.

Internationally Copyright Provides Differing Degrees of Protection to Databases

Copyright law differs internationally in terms of how much protection it extends to factual databases. In the U.S., copyright protects the creative selection and arrangement of data, not the

[4] International differences are discussed later.
[5] U.S. Supreme Court, 499 US 340, 1991.
[6] Note that Web content, such as news articles, music, video, and such, are not data and are protected by copyright law. The focus of this article is on data – such as, in the previous example, the list of items for sale on eBay and their auction prices.

data itself. In other words, the creative choice of what to be included in a database and the creative design of the database schema are protected by copyright in the U.S., but not the factual data records.

Although the U.S. has rejected the "sweat of the brow" doctrine, Australia embraces the doctrine for its copyright law as evidenced by the appeal case *Desktop Marketing Systems Pty Ltd v. Telstra Corporation Limited*[7]. Desktop used all the entries in Telstra's white pages and yellow pages to make CD-ROMs with several additional search features. The Full Court ruled that originality "does not require novelty, inventiveness or creativity", and a work is original "if the compiler has undertaking substantial labour or incurred substantial expense in collecting the information recorded in the compilation." The High Court of Australia confirmed the judgment in 2003 and maintained that Desktop infringed Teslstra's copyright.

The different creativity requirements of the U.S. and Australia represent two extremes. The Canadian law is somewhere in between the extremes. In the judgment of a Canadian case[8], the Court decided that originality "need not be creative, in the sense of being novel or unique." A work is original if it is "more than a mere copy of another work" and requires "an exercise of skill and judgment" that "must not be so trivial that it would be characterized as a purely mechanical exercise."

Despite these differences in the criteria for testing originality, copyright law is quite uniform internationally that one cannot claim copyright protection for individual entries of facts stored in a database.

[7] Full Federal Court of Australia, 2002.
[8] Supreme Court of Canada, CCH Canadian Ltd. V. Law Society of Upper Canada, 2004.

History of Database Legislation

Database creators have tried several ways to protect their non-copyrightable contents[9]. A commonly practiced method is through access control, which often requires user subscription and authentication. But this does not prevent data extraction if the user provides identification to the aggregator (e.g., a user provides login credentials to a financial account aggregator for it to gather information from disparate accounts on the user's behalf [8].) Enforceable contracts to restrict the extraction and reuse of the data are difficult to establish on the Web unless cumbersome "click-through" agreements are in place. As a result, some database creators feel existing law does not give them sufficient protection to their data and their investment in creating databases. Consequently, they have sought means to protect their data through new legislation. See Figure 1 earlier for a summary of legislative activities.

The EU first introduced the Database Directive in 1996 to provide two kinds of protection for a database: copyright for the selection or arrangement of database contents, and *sui generis*[10] right for the contents in the database. The *sui generis* right is a new type of right to prevent unauthorized extraction and/or reutilization of the whole, a substantial part, or systematic extraction and/or reutilization of an insubstantial part, of contents of a database that is created with substantial expenditure. Lawful users are restricted not to "perform acts which conflict with normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database." Here "the legitimate interests" can be broadly interpreted and may not be limited to commercial interests.

[9] Due to limitations on length, we will not discuss all the technical methods that have been used, such as blocking requests from IP addresses that appear to be extracting large quantities of data, etc. In general, for each technical approach to prevent data extraction, there is a possible technical counter-measure to overcome it.
[10] In Latin, meaning "of its own kind", "unique".

The Directive has been criticized for its ambiguity about the minimal level of investment required to qualify for protection [5], its lack of compulsory license provisions [1], the potential of providing perpetual protection under its provision of automatic right renewal after a substantial database update, and the ambiguity in what constitutes a "substantial" update.

Under its reciprocity provision, databases from countries that do not offer similar protection to databases created by EU nationals are not protected by the Directive within the EU. In response, the U.S. database industry pushed the Congress to provide similar protection to database contents. Since then, the Congress has considered six proposals, all of which have failed to pass into law.

HR 3531 of 1996 closely followed the EU Database Directive approach with even more stringent restrictions on data reuse. One of the main concerns is the constitutionality of the scope and strength of the kind of protection for database contents [1,7].

All subsequent U.S. proposals took a misappropriation approach where the commercial value of databases is explicitly considered. HR 2562 of 1998 and its successor HR 354 of 1999 penalize the commercial reutilization of a substantial part of a database if the reutilization causes harm in the primary or any intended market of the database creator. The protection afforded by these proposals can be expansive when "intended market" is interpreted broadly by the creator. At the other end of the spectrum, HR 1858 of 1999 only prevents someone from duplicating a database and selling the duplicate in competition.

Following the reasoning in the NBA v. Motorola case[11], HR 3261 of 2003 has provisions that lie in between the extremes of previous proposals. It makes a data reuser liable for "making available in commerce" a substantial part of another person's database if "(1) the database was

[11] 105 F.3d 841 (2nd Circuit, 1997). Motorola transcribed NBA playoff scores from broadcast and sent them to its pager subscribers. The misappropriation claim by NBA was dismissed.

generated, gathered, or maintained through a substantial expenditure of financial resources or time; (2) the unauthorized making available in commerce occurs in a time sensitive manner and inflicts injury on the database or a product or service offering access to multiple databases; and (3) the ability of other parties to free ride on the efforts of the plaintiff would so reduce the incentive to produce the product or service that its existence or quality would be substantially threatened". The term ''inflicts an injury'' means "serving as a functional equivalent in the same market as the database in a manner that causes the displacement, or the disruption of the sources, of sales, licenses, advertising, or other revenue".

The purpose of HR 3872 is to prevent misappropriation while ensuring adequate access to factual information. It disallows only the free-riding that endangers the existence or the quality of the creator database. Unlike in HR 3261, injury in the form of decreased revenue alone is not an offence.

On December 12, 2005, the Commission of European Communities [2] issued its first evaluation of the Database Directive. The evaluation shows that although the Directive helped harmonize copyright laws within the EU, the economic impact of the *sui generis* right on database production within the EU is unproven. In addition, the scope of the *sui generis* right has proved to the difficult to interpret and its related provisions have "caused considerable legal uncertainty, both at the EU and national level".

These world-wide legislative initiatives demonstrate the substantial difficulties in formulating a database protection law that balances creator incentives and the values added by data reuses. Some of the challenges are briefly discussed below.

Concerns of Providing Legal Protection for Database Contents

Data monopoly. There are situations where data can only come from a sole source due to economy of scale in database creation or impossibility of duplicating the event that generates the data set. For example, no one else but eBay can generate the bidding data of items auctioned on eBay. A law that prevents others from using the factual data from a sole source in effect legalizes a data monopoly which would endanger any downstream value-creating reutilizations of the data. The European Court of Justice (ECJ) partially addressed this issue by trying to distinguish *data created* from *data obtained*, and by protecting only databases whose data is obtained by collecting existing independent materials[12].

Cost distortion. Both the EU database directive and the latest U.S. proposals require substantial expenditure in creating the database for it to be qualified for protection. Database creators thus may over invest at an inefficient level to qualify [10]; see [12] for an economic model that explains such cost distortion.

Update distortion and eternal protection. This is an issue in EU law, which allows for automatic renewal of *sui generis* right when the database has been substantially updated. Such a provision can induce socially inefficient updates solely to attain eternal rights [6].

Constitutionality. Although the U.S. Congress is empowered by the Constitution to regulate interstate commerce under the Commerce Clause[13] and the misappropriation approach often gives a database law a commercial guise, this must be balanced against the Intellectual

[12] European Court of Justice, Grand Chamber, The British Horseracing Board Ltd and Others v. William Hill Organization Ltd., 2004. A database creator with data that is *created*, e.g., BHB, which created the fixture list, would be a natural monopoly if legal protection was granted. Data that is *obtained* presumably could be obtained by anyone willing to make the effort.

[13] Constitution 1.8.3, "To regulate Commerce with foreign Nations, and among the several States, and with the Indian Tribes".

Property Clause[14] which restricts the grant of exclusive rights in intangibles that diminishes access to public domain and imposes significant costs on consumers [4]. Certain database contents are factual data in the public domain; disallowing mere extraction of such data for value-creating activities runs afoul of the very purpose of the Intellectual Property Clause to "promote the Progress of Science and useful Arts". Excessive restrictions on reuse of factual data (a form of speech or press) may also violate the Constitution's First Amendment [3], which protects the freedom of speech and press. Since little extra value for the society as a whole is created by simply duplicating a database in its entirety, preventing verbatim copying of a database is clearly constitutional. A constitutional database law needs to determine how much one is allowed to extract database contents. The constitutional line-drawing between extraction and duplication in data reuse is very difficult [4].

International harmonization. Given the global reach of the Web and increasing international trade, it is desirable to have a harmonized data reuse policy across jurisdictions worldwide. We have discussed some of the differences in the U.S., the EU, Australia, and Canada. A World Intellectual Property Organization (WIPO) study [11] also reveals different opinions from other countries and regions.

A key element to solving these challenges hinges upon finding the right factors for a reasonable balance between protection of incentives and promotion of value creation through data reuse. With this balance, value creation through data reuse is maximally allowed to the extent that the creators still have enough incentives to create the databases. Consensus can develop for international harmonization if we can determine the policy choices that effectively

[14] Constitution 1.8.8, "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries".

balance these factors; a database policy so formulated should survive the scrutiny of constitutionality and other inefficiencies can be avoided or mitigated.

Achieving Balance in Database Legislation

We approach the challenge with an economic model [12] that considers the commercial value of databases. Based on differentiated competition theory, the model considers a database creator, which incurs a cost to create the initial database, and a data reuser, which extracts a certain amount of data from the creator database to create the reuser database. The reuser database can be differentiated from the creator database in terms of scope (e.g., extracting a fraction of the creator's data, combining it with data from other sources) and functionality (e.g., different kind of search algorithm). The reuser uses technology to allow it to easily extract and combine data from existing databases so that the cost of creating the reuser database can be negligible.

The competition from the reuser database can reduce the creator's revenue. When the reduction is such that the creator's revenue cannot offset its cost of creating the database, the market fails[15]. From an economic point view, regulation for data reutilization is needed to prevent or correct market failure.

A regulation potentially can restrict certain stakeholders and benefit certain other stakeholders, but the society as a whole should better off with the regulation. Our analysis shows that such choices depend on the relationship among several factors. The most important two are: (1) the cost of creating the initial database and (2) the level of differentiation between the creator database and the reuser database. The choices[16] in relation to these two factors are depicted in

[15] Market failure is an economic term for the situation where goods or services cannot be provided to consumers (e.g., it is not profitable for creator to produce the database.) Policy intervention can sometimes restore a failed market.

[16] There are actually more than three regions in our paper [12], we have simplified the situation slightly to shorten this paper.

Figure 2, which, as we mentioned earlier "one size does not fit all," illustrates that the policy choices are not just binary.

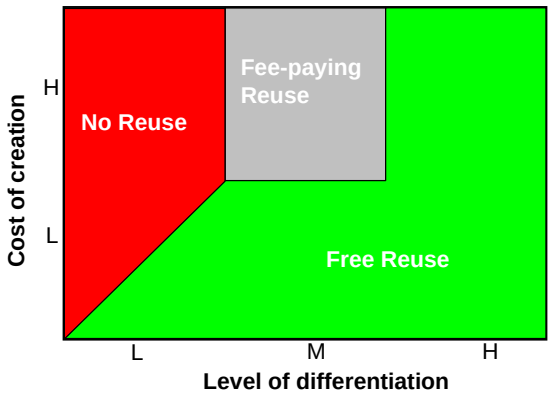


Figure 2. Policy Choices Suggested by the Economic Model

No reuse region. When the level of differentiation is low, not allowing reuse is a reasonable policy choice since such reuse adds little value, and, at the same time, the intense competition can drive the price so low that the creator cannot have enough revenue to offset the cost. Verbatim copying of an entire database is a typical example of this scenario.

Free reuse region. When the level of differentiation is moderate or high, there are two scenarios where free reuse should be allowed: creation cost is low, or differentiation is high regardless of creation cost. With moderate differentiation, competition is not as intense as that in the case of low differentiation. The softened competition allows the creator to make enough revenue to offset its cost. With high differentiation, there will be little competition between the creator and the reuser. In other words, the data reutilization has little impact on the creator.

Although in both cases the reuser could be required to pay the creator a fee, this is not needed to prevent market failure and this is not desirable because there is always an inefficiency associated with money transfer, which is known as transaction cost. The fee can benefit the creator, but it does not create any extra value and the society as a whole incurs a transaction cost.

Fee-paying reuse region. When the level of differentiation is moderate but the cost of creation is high, the reuser should pay a fee to the creator. This is the case where without a fee the reuse would cause market failure, but with a fee the creator can sustain. Since the creator may not be willing to license its data to the reuser, a compulsory licensing provision should be in place.

Some Examples Illustrating the Application of these Principles

The economic model provides a useful framework for facilitating the ongoing debate of database legislation, analyzing data reuse cases, and interpreting court decisions. We will illustrate the applications of the model by revisiting the two cases mentioned earlier.

eBay v. Bidder's Edge. According to our analysis, we need to at least examine the level of differentiation of the database developed by the reuser Bidder's Edge. In terms of searching of bidding data, the reuser database has a much broader coverage; thus, there is competition from the reuser database. In terms of functionality, eBay's database allows one to buy and sell items; the reuser database does not provide any actual auction service. Thus the two databases exhibit significant differentiation. Searching alone does not, in general, reduce eBay's revenue from its auction service. eBay can still compete in the search space, but according to the model eBay should not be given the right to prevent innovative firms such as Bidder's Edge from offering search function before eBay acquires the necessary technical and business skills. Furthermore, if we subscribe to the spin-off theory [5], the eBay database will not meet the cost criterion. Therefore, free reuse by Bidder's Edge should be allowed.

BHB v. William Hill. The ECJ determined that although William Hill did systematically extract and reuse an insubstantial part of BHB's database, the cumulative effect has no possibility for William Hill to "reconstitute and make available to the public the whole and

substantial part of the contents of the BHB database" and therefore "seriously prejudice the investment" in the creation of the database. The criterion of "reconstitution" effect can be explained using the economic model as the reuser database having little differentiation. The ECJ also stressed that the injury needs to be serious, which can be understood from the market failure perspective in the model.

BHB spends £4 million annually to maintain the database. The ECJ judgment provides a guideline for determining if this cost is protected by the Database Directive. After making the distinction between creating and obtaining data, the ECJ determined that the investment protected by the *sui generis* right "does not cover the resources used for creating the materials which make up the contents of a database." To create the racing list, BHB had to verify information of participants, e.g., a horse's age and pedigree, and such information was *obtained* by BHB. The ECJ further ruled that "The resources used for verification during the stage of creation of materials" are not part of protected investment. These cost accounting rules used by the ECJ constitute a particular standard of determining the cost factor in the model.

Conclusion

Although the legislative efforts may seem to have stalled in the U.S. during the past two years, the issues related to technology-enabled data reuse have not been resolved. We discussed these issues and presented the preliminary results of an economic analysis on how to balance the benefits of data reuse to society and the interests of profiting from creating the initial databases[17]. The results show there is not a one-size-fits-all formula for data reuse regulation. Rather, depending on several factors, no reuse, free reuse, or fee-paying reuse are welfare-enhancing choices.

[17] There are many other factors, such as the political, legal, and enforcement processes in different jurisdictions, that are beyond the scope of this paper. The intention of this paper is to establish some basic principles that could facilitate these other processes.

As technologies for reusing data from various sources continue to emerge and improve, the need for understanding the legal implications of applying these technologies will become increasingly acute. We are continuing to develop further understanding of the issues related to applying data reuse technologies. We anticipate the research to bring us closer to finding the right balance with which serendipitous and innovative data reutilization can be maximally allowed to provide value-added services without diminishing the incentives of compiling databases and making them available on the web.

References

1. Colsten, C. Sui Generis Database Right: Ripe for Review? *The Journal of Information, Law and Technology* 3 (2001).
2. Commission of the European Communities (CEC). First Evaluation of Directive 96/9/EC on the Legal Protection of Databases. December 12, 2005, Brussels.
3. Grove, J. Wanted: Public Policies That Foster Creation of Knowledge. *Communications of the ACM* 47, 5 (2004), 23-25.
4. Heald, P.J. The Extraction/Duplication Dichotomy: Constitutional Line Drawing in the Database Debate. *Ohio State Law Journal* 62, 2 (2001) 933-944.
5. Hugenholtz, P.B. Program Schedules, Event Data and Telephone Subscriber Listings under the Database Directive: The "Spin-Off" Doctrine in the Netherlands and elsewhere in Europe. 11th Annual Conference on International Law & Policy (2003), New York.
6. Koboldt, C. The EU-Directive on the legal protection of databases and the incentives to update: An economic analysis. *International Review of Law and Economics* 17, 1 (1997) 127-138.
7. Lipton, J. Private Rights and Public Policies: Reconceptualizing Property in Databases. *Berkeley Technology Law Journal* 18, 3 (2003) 773-852.
8. Madnick, S.E., Siegel, M. D. Seize the Opportunity: Exploiting Web Aggregation. *MISQ Executive* 1, 1 (2002) 35-46.
9. O'Rourke, M.A. Is Virtual Trespass an Apt Analogy? *Communications of the ACM*, 44, 2 (2001), 98-103.
10. Samuelson, P. Legal Protection of Database Contents. *Communications of the ACM* 39, 12 (1996), 17-23.
11. Tabuchi, H. International Protection of Non-Original Databases: Studies on the Economic Impact of the Intellectual Property Protection of Non-Original Databases. CODATA (2002), Montreal, Canada. http://www.codata.org/codata02/03invited/Tabuchi/Tabuchi_CODATA_ejournal.pdf.
12. Zhu, H., Madnick, S.E., Siegel, M.D. Policy for the Protection and Reuse of Non-Copyrightable Database Contents. MIT Sloan School Working Paper (2005) #4751-05. Available at SSRN: http://ssrn.com/abstract=876960.

We have been discussing the **online travel industry** <http://sramanamitra.com/blog/929> and have covered **Yahoo! Travel** <http://www.readwriteweb.com/archives/yahoo_travel_he_1.php> from a **Web 3.0** <http://sramanamitra.com/blog/572> perspective already. Here we take a look at the offering from online travel community behemoth **TripAdvisor** <http://www.tripadvisor.com> .

TripAdvisor was founded in February 2000 and is among the worldÄôs largest online travel communities with over 20 million unique monthly visitors and approximately 5 million registered members. TripAdvisor is currently part of **Expedia** <http://www.expedia.com> (**NASDAQ** <http://www.nasdaq.com> : EXPE). The site is a winner of **PC Magazine's** <http://www.pcmag.com> Top 100 Web Sites and **Forbes'** <http://www.forbes.com> Best of the Web.

Context

TripAdvisor provides recommendations for hotels, resorts, inns, vacation packages, and travel guides. The site is broken up into distinct categories like Find Hotels, Flights, Read & Write Reviews, Browse Destinations, Rants & Raves, GoLists, TripAdvisor Forums, Helpful Links, Top Business Hotels, TripAdvisor Inside, and Photo and Video Sharing, but the organization could be better. The

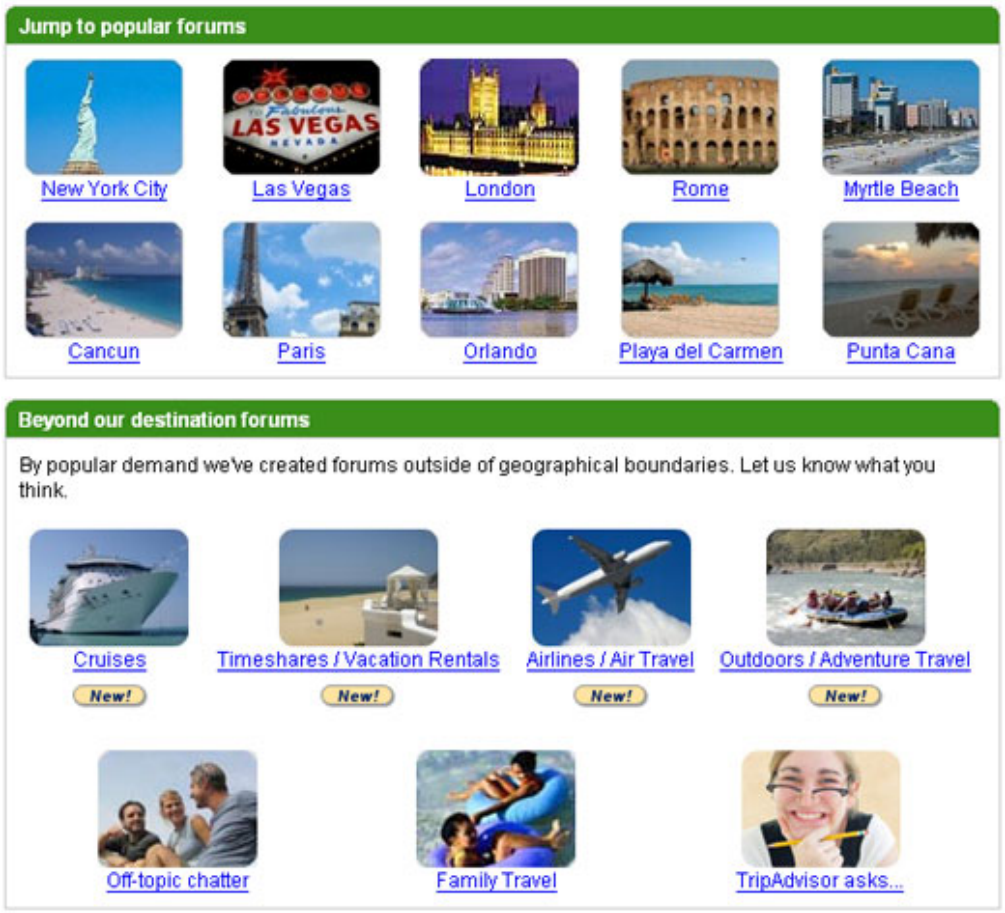
users have the freedom of moving quickly from one category to another, but the organization doesn't necessarily create an integrated contextual experience. TripAdvisor doesn't flow with the natural rhythm of the travel planning experience.

In fact, the key problem with TripAdvisor is its organization. I tried to look at the photos of Giraffe Manor B&B in Nairobi, but after scrolling through numerous pages, I couldn't find them, even though the reviewer claims to have posted them.

Content

TripAdvisor has a wide range of content. The site contains information on over 180,000 hotels and 91,000 restaurants in 23,000 cities. Users can also browse travel destinations across the world with the aid of a travel map and as the searches narrow down, the user is provided with a local map showing local attractions and the best deals for local hotels.

TripAdvisor is wiki-enabled, which facilitates millions of travelers to view, contribute, and edit the guides available on more than 24,000 destinations worldwide. The site also has photos and videos. The site has tie-ins with over 17 business partners in the travel industry including Expedia, **Sabre** <http://www.sabre.com> , **Orbitz** <http://www.orbitz.com> and **American Airlines** <http://www.aa.com> .



Community

TripAdvisor has the largest travel community on the web, which is visited by more than 500,000 travelers every day. The **TripAdvisor Forum** <http://www.tripadvisor.com/ForumHome> allows users to post their experiences about tours, express opinions, recommend hotels, resorts, inns, vacations, travel packages, vacation packages, post questions and answer or advise other members of the forum. Users can also post photos and videos of their tours. TripAdvisor allows users to create a travel blog on **TravelPod** <http://www.travelpod.com> .

TripAdvisor is by far the most successful in engaging a global community of travelers in sharing their experiences and reviews on the site. The Rant & Rave function can make or break the reputation of a hotel or a restaurant in a nanosecond, and is tremendously helpful to travelers!

Commerce

TripAdvisor has tie-ins with a number of commerce sites such as its parent Expedia, **Hotels.com** <http://www.hotels.com> , **British Airways** <http://www.britishairways.com> , **Delta** <http://www.delta.com> ,

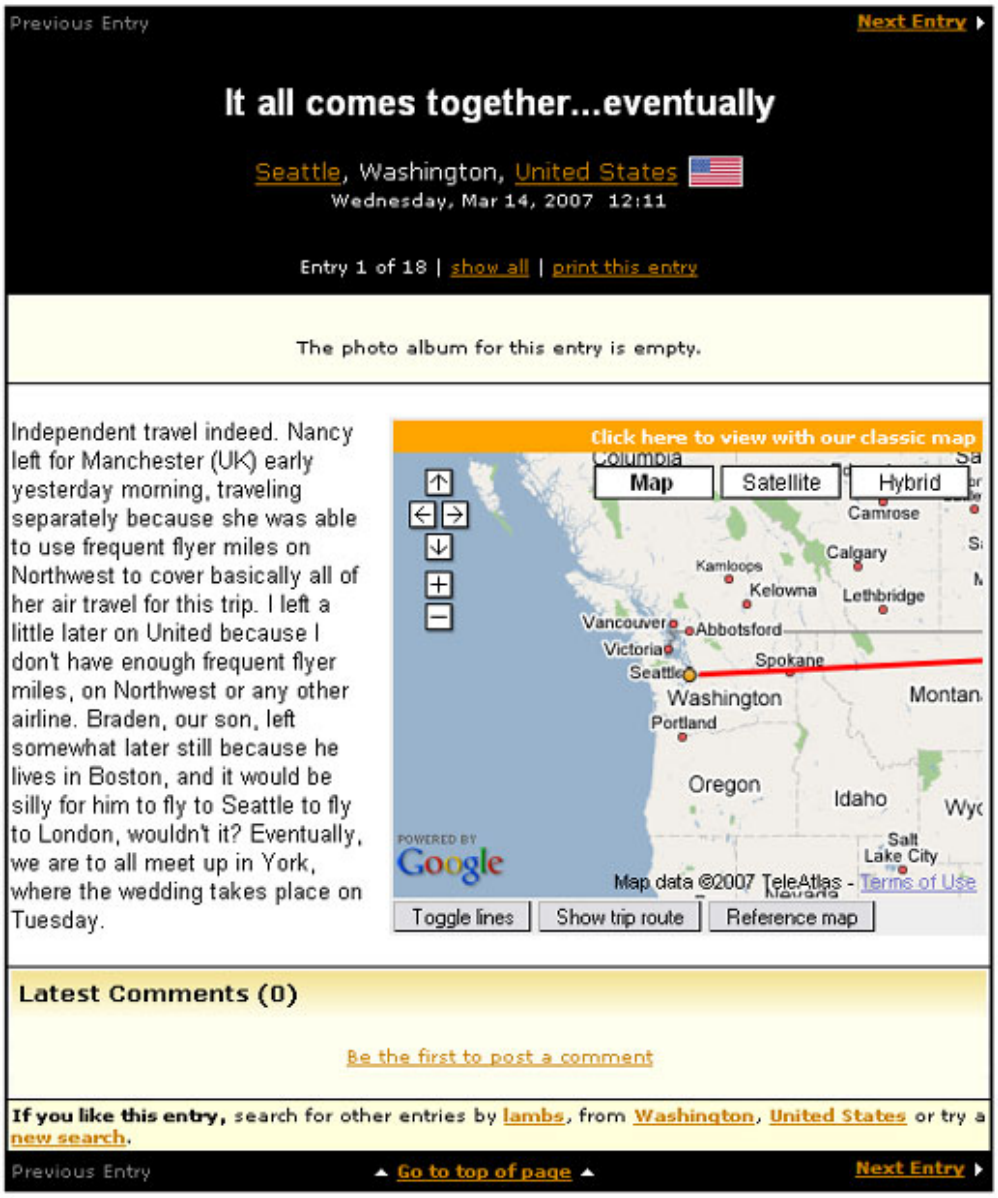
Priceline <http://www.priceline.com> and **Lastminute.com** <http://www.lastminute.com> , all of which aid its users in booking flights, hotels, vacations or cruises, enabling the site to earn commission revenue. This, however, is a commodity function, available on all travel sites.

The TripAdvisor Store retails various **TripAdvisor Gear** <http://www.cafepress.com/tripadvisor> through a partnership with **Cafepress** <http://www.cafepress.com> . Items sold by TripAdvisor include hats, mugs, clothing, bags etc. The site has identified a way of monetizing its brand, but so far, this looks like a fairly shabby effort, since to be blunt, the merchandising, by and large, sucks. They should look at how National Geographic does its merchandising, by creating unique products sourced from various parts of the world - jackets from Nepal, wool slippers from Tibet, caps from Peru - rather than this bland catalog of insignia products.

Personalization

TripAdvisor offers some good personalization and travel planning options. Each personalized page contains full information about the user, stating the personÄôs recent travels or booking, contributions to TripAdvisor, reviews and also includes user preferences for travel (pleasure or business), spending habits, and vacation choices. The personalization facility allows users to organize and plan oneÄôs trip, save hotels, attractions, compare hotels, make a list of places one would like to visit, add maps and notes, organize items by destination or days, and create a personal travel guidebook to save, print or email.

The site also informs the registered users with a time-sensitive e-mail newsletter for travelers planning a vacation, giving customized e-mail alerts on specific hotels, attractions and cities of their choice. The site also has other personalized newsletters like TripWatch and Weekend Getaway Guide provided through email.



Vertical Search

TripAdvisor offers user-friendly search options for hotels and flights enabling users to select from multiple options according to their preferences, but there is nothing special or overly different about it.

I would like to plan a trip centered around B&Bs in Andalucia (Southern Spain). How do I do that? The vertical search option simply doesn't get sophisticated enough quite yet.

Business Model

TripAdvisor has an Alexa traffic rank of 504 and has more than 20 million

unique monthly visitors. The site has display advertising as well as cost-per-click advertising. **Travel Ad Network** <http://traveladnetwork.com> is TripAdvisorÄôs exclusive advertising representative for display advertising. Advertising and Commissions on bookings constitute their primary revenue streams.

Conclusion

My final **Web 3.0 Rating** <http://sramanamitra.com/blog/572> is: Context: A-; Content: A-; Community: A+; Commerce: B-; Personalization: B+; Vertical Search: B-; Overall : A-

#ANALYSIS <HTTP://WWW.READWRITE.COM/TAG/ANALYSIS>

#WEB <HTTP://WWW.READWRITE.COM/TAG/WEB>

THE WALL STREET JOURNAL.
WSJ.com

WEEKEND JOURNAL | June 1, 2007

Deconstructing TripAdvisor

Nancy Keates on how seasoned travelers decode the ratings on the nation's most influential hotel review site.

By NANCY KEATES

'Simply the best!!!!!!" "Very Grand!!" "Awesome." "Unbelievable."

The last one might be the most accurate.

You'd think a reporter who has covered travel for this newspaper for more than a decade wouldn't be fooled by such superlatives. But on a trip in January my family discovered that some people who write reviews on TripAdvisor.com are thrilled to pay $280 to spend the night next to an eight-lane highway. Ranked on the Web site as the No. 1 property in Carlsbad, Calif. -- ahead of the Four Seasons Aviara and the famous La Costa spa -- the West Inn & Suites wasn't only far from the center of the quaint oceanside town, it was also next to a working train track with a view of a large power plant.

A Second Opinion

Many experienced travelers compare TripAdvisor reviews against those on major travel booking sites. Here are some other sites with user-generated hotel reviews.

For many, TripAdvisor has become a first stop for travel planning. Thanks in part to its prominence in Google searches, some 24 million visitors a month check out what other users have to say about where to stay, eat and play around the world. (In contrast, publisher Frommer's sells 2.5 million guidebooks a year.) With more than 250,000 hotels, its sheer breadth of properties makes it more useful than other hotel Web sites. Its wide range of contributors -- there are nearly 10 million reviews and opinions -- make it more democratic. At a conference in November, the chief executive of guidebook publisher Lonely Planet said the Web site's influence is so great that the company considered eliminating hotel reviews altogether. TripAdvisor is also gobbling up a number of other sites that do things like search for low airfares and list seating charts on airplanes.

But relying on the wisdom of crowds can be dangerous. When Chirag Chotalia, a private-equity investor from New York, booked a long weekend at the Ritz-Carlton South Beach in Miami this March, he was swayed by two reviews. One raved about the "excellent" staff; the other called the service "stellar." Instead, the 25-year-old says he found surly, unprofessional concierges, a long wait at check in and an under-staffed pool. A spokeswoman says that the "overwhelming majority" of guests are very happy with their stays.

Luxurious in Liverpool

To avoid such pitfalls, it is necessary to deconstruct every review -- and its author. After all, a hotel recommended by a once-a-year vacationer could be a disaster for a business traveler. What someone from Liverpool, England, finds luxurious might not appeal to a picky Manhattanite. Roosters crowing at dawn may not seem worth mentioning to some reviewers, while others might think harping on things like an "intermittent electrical buzzing in the air unit" (see Amalfi Hotel, Chicago) is a tad excessive.

In an attempt to decode TripAdvisor, I interviewed heavy users and spoke to online-travel experts. The most common, and most obvious, place to start when determining reliability is to weed out reviews that are way off the mean: those that have one star when the rest are positive, or five stars when the others are mixed. That indicates either an unusual incident or a writer with some interest in the hotel, like a rival property or the general manager's friend. Other hints a review might be fake: The writer mentions a nearby property as superior, has only written about that one hotel and has only visited the site once -- on the day of the review. (You can check for other hotels a writer has evaluated by clicking on the reviewer's name.)

David Brinley

Next, study the reviewer as closely as the review. In February, Juan Padro, a headhunter from North Grafton, Mass., was weighing a trip to Ladera in Soufriere, St. Lucia -- a resort that elicited mixed opinions on TripAdvisor. Some guests raved about the privacy, peace and beauty, while others complained it was too much like a campground to justify the average $990-a-night rate.

Mr. Padro didn't make up his mind until he read a review entitled "What a rip off!" It said, "one of the things that MUST BE MENTIONED is the fact that the moment the sun starts going down, the beautiful chorrus [sic] of frogs starts their singing until the sun rises again. It was really hard to fall asleep with all that noise." The reviewer complained that the music and the atmosphere in the bar was "VEEEEERY RELAXED" -- and left for South Beach, Miami, two days into a 10-day stay. "The guy was so clearly a meathead," says Mr. Padro, who decided (correctly, it turned out) that any resort that would scare off someone like that would be perfect for him.

Excessive effusiveness is a red flag for Wayne Rutman, a private investor from Wilmington, Del., who is on the road every month and frequently uses TripAdvisor to plan his trips. Phrases like "dream vacation of a lifetime" and "best place I ever stayed" signal a lack of experience. People who find it necessary to say they're world travelers in the first line are also suspicious, like someone who feels the need to impress others at a cocktail party, he says.

Some Good Finds

Often the reviews are dead-on. Fusion Suites, the bed and breakfast ranked as the No. 1 property in Amsterdam, is an amazing find, with enormous rooms located on a tree-lined street near the Van Gogh museum. I had never heard of Eastgate Tower near the United Nations in New York when I took a chance (again) on Trip-Advisor and booked it for a recent family vacation. The $250-a-night suite had two large bedrooms with two beds in each, two bathrooms, a living room and a full kitchen; it was clean and well-staffed; there was even a bellman who carried bags.

Where a reviewer lives -- a detail listed right next to the user name -- can be a telling clue. Among heavy users in the U.S., there are ongoing discussions about whether non-Americans can be trusted. "Europeans have different standards," says Loren Medina, a school social worker in

Paramus, N.J., who travels with her husband and children. "The rooms are smaller, they're in older buildings with older plumbing. They find more things acceptable."

Mr. Chotalia, the New York investor who was disappointed with his stay at the Ritz-Carlton South Beach, wonders if such geographic issues played a role. The upbeat reviewers "sounded educated and worldly so I thought I could respect their opinions," he says. But when he looked again, he noticed the first was from Manchester, England, and the second from Canada. ("There are cultural differences between Canadians and Americans," says a spokesman for the Canadian embassy in Washington.)

The reviewer's hometown can count even within the U.S. Bob McDevitt (whose screen name is Cap10Bob), doesn't believe anything written by a New Yorker. The 58-year-old salesman from Boston says "people from there wouldn't like anything anyway." Here's an excerpt from Mr. McDevitt's TripAdvisor write-up of the Westin Rio Mar in Puerto Rico (which is now the Rio Mar Beach Resort & Spa, a Wyndham Grand Resort): "A group of five middle age golfers/fishermen/general tourists, stayed for 4 days in mid-March. We found the hotel to be excellent."

One TripAdvisor reviewer complained about noisy frogs at the Ladera in St. Lucia (top). The site ranks West Inn & Suites (above left) No. 1 in Carlsbad, Calif., and Eastgate Tower (above right) No. 49 in New York City.

New York banker Aylin Ural, 35, wrote a review of the same hotel a week later. "Every morning starting at 6 a.m. we awoke to people walking above us, doors slamming constantly, toilets flushing incessantly, and people from the parking lot shouting. This is EVERY morning. We are from Manhattan so we are used to noise." She says reviews by Manhattanites are often the only ones she'll believe. "We have certain standards," she says. Many accolades by her brethren ("beautiful beaches" and "a true paradise") steered her to spend her honeymoon at Marriott Frenchman's Reef in St. Thomas despite numerous TripAdvisor dissenters who hailed from other locales. She loved it.

The reviewer's user name matters to Michelle Hill, who lives in Lake Placid, N.Y., and travels with her husband and kids several times a year. A recent report by someone called "crzy4cncun" was believable, she decided, because it meant that person had a lot of experience in Cancun. In the review, crzy4cncun mentioned that she had a teenager: That was an added bonus for Ms. Hill, since finding somewhere kids that age can be happy is difficult, she says.

Ms. Hill always clicks on a user name to check what other properties that person has written about. She considers herself "very particular" -- a Westin Hotels kind of person as opposed to a Best Western gal. Review writers who stay in chains she would avoid wouldn't understand what she's looking for, she says. The jackpot: Finding someone who has reviewed a property where she's also stayed so she knows if they're in sync.

In general, TripAdvisor is more helpful for smaller, more obscure properties that aren't fully covered by other sources. It can be time-consuming and less effective for well-known hotels because they have so many reviews that are often so widely disparate, making it hard to get a sense of the property. Orlando, Fla.-based consultant Mark Feinberg discovered that when he was planning a trip for this month to New York to celebrate his daughter's 12th birthday. Finding himself stuck on the Web site for hours trying to decide between the Four Seasons and the Ritz-Carlton, he finally went with his cousin's advice and chose the Four Seasons because he was so confused by what he read on TripAdvisor.

The Four Seasons reviews ranged from "Wow, what a place" to "Nightmare after nightmare." According to the latter: "There was still feces flecking the toilet when we checked in and a hair on the nice, white sheets...The front-desk staff were gruff and unhelpful -- failing to even answer basic questions about museums and theatre tickets." Comments on the Ritz-Carlton Central Park also ran the gamut, from "Missed it by very much" to "Perfect Stay."

The Four Seasons hotel's director of marketing, Brian Honan, says the chain takes feedback "really seriously," and that the hotel has no record of any such complaints over the dates the guest stayed at the hotel. A Ritz-Carlton spokeswoman says it views comments from guests as a chance to "continue to improve."

Up-to-Date Details

When certain key words ("hurricane" or "construction") pop up, TripAdvisor is at its best. It is one of the few places to find indications that a recent event has affected the hotel's quality. Mr. Feinberg learned that lesson the hard way when he stayed at the Renaissance Resort at the World Golf Village in Saint Augustine, Fla., a few months after a hurricane. Reviews he read in golfing magazines had raved about the place, but he smelled mildew and mold everywhere. When he looked on TripAdvisor afterwards, he saw people had mentioned the problem.

Renaissance General Manager Mark Schwantner says the resort did experience a problem after the hurricane knocked out power for a few days. Since then the resort has spent over $1 million adding new dehumidification units and resealing the building; it now monitors interior humidity levels to make sure they don't exceed 55%.

When TripAdvisor started in 2000, the site was a search engine that hooked into travel information already on the Web -- from newspapers, magazines, online guidebooks, chat rooms, message boards and personal home pages. As traffic grew, people started adding their own reviews, which soon became the most-read pages on the site.

"When we first thought of pushing the user reviews, we were actually a little nervous about whether the site would just turn into a gripe site," says TripAdvisor co-founder and chief executive officer Stephen Kaufer. Instead, most of the reviews were overwhelmingly positive. That gave the company the idea to earn money by "contextual-commerce links," allowing consumers to make a reservation through links to booking sites.

Focusing on Transactions

Hotel-booking sites started to see the value in that. In 2004, IAC/InterActiveCorp bought TripAdvisor for an estimated $430 million and wrapped it into its Expedia group. (It later spun off the Expedia group, including TripAdvisor, into a separate company.) Revenue from the 173-person company comes from travel-related advertising and the fees TripAdvisor gets from online-booking sites when users click to make a reservation. It often isn't enough just to have a lot of traffic on a Web site, says Scott Kessler, an equity analyst with Standard & Poor's. Since TripAdvisor has such high-quality traffic (people who use it have a great interest in making a purchase since they are considering a trip) it makes financial sense to take that traffic and try to turn it into revenue, says Mr. Kessler.

As a result, TripAdvisor has shifted from solely a forum-like site to more of a transaction-based model. In August 2006, the company changed its format so that instead of going directly to hotel reviews, the home page's default became similar to what you'd find on airline-booking Web sites. Consumers enter dates of travel and destination and are presented with a list of properties they can book online. That's a different list from the top hotels as ranked by TripAdvisor users; to get to that page, users have to type the name of the city followed by the word "hotels" in the main search bar.

Sometimes the drive to monetize can be at odds with the drive to be consumer friendly. TripAdvisor doesn't give a hotel's Web address unless that hotel pays it to do so, encouraging visitors to use online-booking sites (including Expedia, Orbitz, Hotels.com and others) and discouraging them from leaving the Trip-Advisor site. "If all we did was look out for consumers, we'd provide a link that would take you to a hotel's Web site," says Mr. Kaufer. "It does absolutely conflict with our interest in making money." He says TripAdvisor looks out for consumers in many other ways and that there are paid links to hotel sites.

The company's revenue is still small, at $105 million in 2006, compared with sites like Expedia and Travelocity. However, with profit margins estimated above 50% and a growth rate thought to be over 50% a year, the site offers potential at a time when hotels and airlines are trying to take back online bookings and get consumers to go directly to their sites, says Aaron Kessler, an analyst at Piper Jaffray Companies.

As the Web site has evolved, so have the users. It is possible to see how other reviewers rate a review, a feature called "Helpful Votes." People can also pick a few hotels off a destination's top rankings and then go into the TripAdvisor forums, where locals tend to respond. That's where the site's addicts often congregate as well. Over 530,000 members have posted to the forums since the site started them three years ago, the company says. Of those, 769 have posted more than a thousand times. The most active member posted 20,593 times.

Taking the time to open photos posted by reviewers helps users get a sense of the writer. One noteworthy example: a picture of a pair of dirty socks used to illustrate a lapse in housekeeping. Frequent TripAdvisor visitors also use the subrankings (including "Romance," "Families" and "Singles") and the information listed to the right of reviews that give the writer's age, purpose of their trip and reasons for selecting the hotel. A new feature on the site lets users email a reviewer directly to get more information.

Though the West Inn is still rated the top hotel in Carlsbad, since my family's stay there reviewers have remarked about the hotel's downsides, including its location. Why didn't that happen earlier? Kim Akers, the hotel's general manager, says people did mention its proximity to the highway (and that I didn't go far back enough into the reviews) but that in most cases they all said it didn't diminish the experience because the hotel has triple-paned windows, music piped in outside by

the pool and a shuttle to take guests to downtown Carlsbad.

Then again, there are some things people just won't tell you. Tara Yelman, a divorce attorney from San Diego, found the Four Seasons Hualalai on the Big Island, Hawaii, through TripAdvisor and asked for a room as quiet as possible after reading some complaints about thin walls. The room she stayed in -- an oceanfront with the best full-ocean view on the property, separated from most of the other rooms at the hotel -- is now so precious to her she won't ever give away the room number. Especially not on TripAdvisor.

A Second Opinion

Another tip: Many experienced travelers compare TripAdvisor reviews against those on major travel booking sites. But there are some smaller options out there. Here are some other sites with user-generated hotel reviews.

SITE	# OF REVIEWS	COMMENT
Fodors.com	125,000	The user-generated reviews are generally more helpful than the Fodor's reviews, which tend to include more information than opinion and rarely anything too negative. Reviews are limited to the restaurants and hotels already in the database -- places a spokeswoman says the company identifies as quality locations. Information about the commenters is limited to their home towns.
Gusto.com	15,000	This independent start-up based in Springfield, Mo., has an audience that's 70% female with an average age of 39 years. It covers hotels around the world but, as founder Jeff Wasson says, it is "North America-centric." Site has links to writer's profiles and hotel Web sites.
HotelShark.com	1,300	Each property on this site, run by an independent Palo Alto, Calif., company, includes a composite summary of reviews in a Zagat-like approach -- but often there's only one review to summarize. That's because the company screens reviews and only accepts ones it finds "sincere," says creator Ken Marshall. Reviews that are no longer applicable are removed.
IgoUgo.com	45,000	Launched in 2000 and owned by Travelocity, IgoUgo's coverage is vast, though the reviews tend to be short and over a year old. ("We're streamlining that process to make it easier to submit," says Peter Campion, general manager.) The site provides a lot of background information on the person writing the review, including their dream destinations and favorite movies and bands.
Zoomandgo.com	35,000	Travelers from all over the world submit reviews and video clips of hotels and vacations, though 70% of users come from the U.S. The most popular destinations are the Caribbean and major U.S. cities, but founder Jonathan Haldane says there are a surprising number of video clips from Hong Kong.

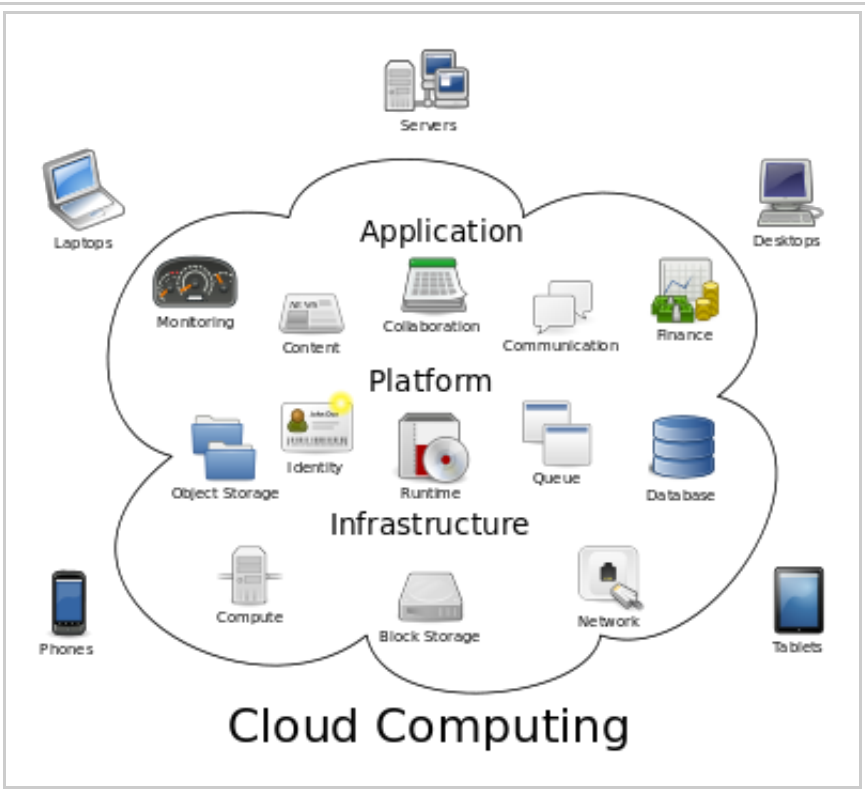
Write to Nancy Keates at nancy.keates@wsj.com

Cloud computing

From Wikipedia, the free encyclopedia

Cloud computing is a colloquial expression used to describe a variety of different computing concepts that involve a large number of computers that are connected through a real-time communication network (typically the Internet).[1] Cloud computing is a jargon term without a commonly accepted non-ambiguous scientific or technical definition. In science, cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time. The popularity of the term can be attributed to its use in marketing to sell hosted services in the sense of application service provisioning that run client server software on a remote location.

Contents



Cloud computing logical diagram

Advantages

Cloud computing relies on sharing of resources to achieve coherence and economies of scale similar to a utility (like the electricity grid) over a network.[2] At the foundation of cloud computing is the broader concept of converged infrastructure and shared services.

The cloud also focuses on maximizing the effectiveness of the shared resources. Cloud resources are usually not only shared by multiple users but as well as dynamically re-allocated as per demand. This can work for allocating resources to users in different time zones. For example, a cloud computer facility which serves European users during European business hours with a specific application (e.g. email) while the same resources are getting reallocated and serve North American users during North America's business hours with another application (e.g. web server). This approach should maximize the use of computing powers thus reducing environmental damage as well, since less power, air conditioning, rackspace, and so on, is required for the same functions.

The term moving cloud also refers to an organization moving away from a traditional capex model (buy the dedicated hardware and depreciate it over a period of time) to the opex model (use a shared cloud infrastructure and pay as you use it)

Proponents claim that cloud computing allows companies to avoid upfront infrastructure costs, and focus on projects that differentiate their businesses instead of infrastructure.[3] Proponents also claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less maintenance, and enables IT to more rapidly adjust resources to meet fluctuating and unpredictable business demand.[3][4][5]

Hosted services

In marketing, cloud computing is mostly used to sell hosted services in the sense of Application Service Provisioning that run client server software on a remote location. Such services are given popular acronyms like 'SaaS' (Software as a Service), 'PaaS' (Platform as a Service). End users access cloud-based applications through a web browser or a light-weight desktop or mobile app while the business software and user's data are stored on servers at a remote location.

History

The 1950s

The underlying concept of cloud computing dates back to the 1950s, when large-scale mainframe computers became available in academia and corporations, accessible via thin clients/terminal computers, often referred to as "dumb terminals", because they were used for communications but had no internal computational capacities. To make more efficient use of costly mainframes, a practice evolved that allowed multiple users to share both the physical access to the computer from multiple terminals as well as to share the CPU time. This eliminated periods of inactivity on the mainframe and allowed for a greater return on the investment. The practice of sharing CPU time on a mainframe became known in the industry as time-sharing.[6]

The 1960's–1990's

John McCarthy opined in the 1960s that "computation may someday be organized as a public utility."[7] Almost all the modern-day characteristics of cloud computing (elastic provision, provided as a utility, online, illusion of infinite supply), the comparison to the electricity industry and the use of public, private, government, and community forms, were thoroughly explored in Douglas Parkhill's 1966 book, *The Challenge of the Computer Utility*. Other scholars have shown that cloud computing's roots go all the way back to the 1950s when scientist Herb Grosch (the author of Grosch's law) postulated that the entire world would operate on dumb terminals powered by about 15 large data centers.[8] Due to the expense of these powerful computers, many corporations and other entities could avail themselves of computing capability through time sharing and several organizations, such as GE's GEISCO, IBM subsidiary The Service Bureau Corporation (SBC, founded in 1957), Tymshare (founded in 1966), National CSS (founded in 1967 and bought by Dun & Bradstreet in 1979), Dial Data (bought by Tymshare in 1968), and Bolt, Beranek and Newman (BBN) marketed time sharing as a commercial venture.

The 1990s

In the 1990s, telecommunications companies, who previously offered primarily dedicated point-to-point data circuits, began offering virtual private network (VPN) services with comparable quality of service, but at a lower cost. By switching traffic as they saw fit to balance server use, they could use overall network bandwidth more

effectively. They began to use the cloud symbol to denote the demarcation point between what the provider was responsible for and what users were responsible for. Cloud computing extends this boundary to cover servers as well as the network infrastructure.[9]

As computers became more prevalent, scientists and technologists explored ways to make large-scale computing power available to more users through time sharing, experimenting with algorithms to provide the optimal use of the infrastructure, platform and applications with prioritized access to the CPU and efficiency for the end users.[10]

Since 2000

After the dot-com bubble, Amazon played a key role in all the development of cloud computing by modernizing their data centers, which, like most computer networks, were using as little as 10% of their capacity at any one time, just to leave room for occasional spikes. Having found that the new cloud architecture resulted in significant internal efficiency improvements whereby small, fast-moving "two-pizza teams" (teams small enough to feed with two pizzas) could add new features faster and more easily, Amazon initiated a new product development effort to provide cloud computing to external customers, and launched Amazon Web Services (AWS) on a utility computing basis in 2006.[11][12]

In early 2008, Eucalyptus became the first open-source, AWS API-compatible platform for deploying private clouds. In early 2008, OpenNebula, enhanced in the RESERVOIR European Commission-funded project, became the first open-source software for deploying private and hybrid clouds, and for the federation of clouds.[13] In the same year, efforts were focused on providing quality of service guarantees (as required by real-time interactive applications) to cloud-based infrastructures, in the framework of the IRMOS European Commission-funded project, resulting to a **real-time cloud environment**.[14] By mid-2008, Gartner saw an opportunity for cloud computing "to shape the relationship among consumers of IT services, those who use IT services and those who sell them"[15] and observed that "organizations are switching from company-owned hardware and software assets to per-use service-based models" so that the "projected shift to computing ... will result in dramatic growth in IT products in some areas and significant reductions in other areas."[16]

On March 1, 2011, IBM announced the IBM SmartCloud framework to support Smarter Planet.[17] Among the various components of the Smarter Computing foundation, cloud computing is a critical piece.

Growth and popularity

The development of the Internet from being document centric via semantic data towards more and more services was described as "Dynamic Web".[18] This contribution focused in particular in the need for better meta-data able to describe not only implementation details but also conceptual details of model-based applications.

The ubiquitous availability of high-capacity networks, low-cost computers and storage devices as well as the widespread adoption of hardware virtualization, service-oriented architecture, autonomic, and utility computing have led to a growth in cloud computing.[19][20][21]

Financials Cloud vendors are experiencing growth rates of 90% per annum.[22]

Origin of the term

The origin of the term *cloud computing* is unclear. The expression *cloud* is commonly used in science to describe a large agglomeration of objects that visually appear from a distance as a cloud and describes any set of things whose details are not inspected further in a given context. The expression somehow can be interpreted as a loosely coherence and moving links between entities in a conformed community of objects.

- Meteorology: a weather cloud is an agglomeration.
- Mathematics: a large number of points in a coordinate system in mathematics is seen as a point cloud;
- Astronomy: many stars that crowd together are seen as star clouds (also known as star mist) in the sky, e.g. the Milky Way;
- Physics: The indeterminate position of electrons around an atomic kernel appears like a cloud to a distant observer;
- Video Games: "The Cloud" was what followed Mario characters around, allowing them to store and access extra items;

In analogy to above usage the word *cloud* was used as a metaphor for the Internet and a standardized cloud-like shape was used to denote a network on telephony schematics and later to depict the Internet in computer network diagrams. The cloud symbol was used to represent the Internet as early as 1994.[23][24] Servers were then shown connected to, but external to, the cloud symbol.

Urban legends claim that usage of the expression is directly derived from the practice of using drawings of stylized clouds to denote networks in diagrams of computing and communications systems.

The term became popular after Amazon.com introduced the Elastic Compute Cloud in 2006.

Similar systems and concepts

Cloud Computing is the result of evolution and adoption of existing technologies and paradigms. The goal of cloud computing is to allow users to take benefit from all of these technologies, without the need for deep knowledge about or expertise with each one of them. The cloud aims to cut costs, and help the users focus on their core business instead of being impeded by IT obstacles.[25]

The main enabling technology for cloud computing is virtualization. Virtualization abstracts the physical infrastructure, which is the most rigid component, and makes it available as a soft component that is easy to use and manage. By doing so, virtualization provides the agility required to speed up IT operations, and reduces cost by increasing infrastructure utilization. On the other hand, autonomic computing automates the process through which the user can provision resources on-demand. By minimizing user involvement, automation speeds up the process and reduces the possibility of human errors.[25]

Users face difficult business problems every day. Cloud computing adopts concepts from Service-oriented Architecture (SOA) that can help the user break these problems into services that can be integrated to provide a solution. Cloud computing provides all of its resources as services, and makes use of the well-established standards and best practices gained in the domain of SOA to allow global and easy access to cloud services in a standardized way.

Cloud computing also leverages concepts from utility computing in order to provide metrics for the services used. Such metrics are at the core of the public cloud pay-per-use models. In addition, measured services are an essential part of the feedback loop in autonomic computing, allowing services to scale on-demand and to perform automatic failure recovery.

Cloud computing is a kind of grid computing; it has evolved from grid computing by addressing the QoS (quality of service) and reliability problems. Cloud computing provides the tools and technologies to build data/compute intensive parallel applications with much more affordable prices compared to traditional parallel computing techniques.[25]

Cloud computing shares characteristics with:

- Client–server model — *Client–server computing* refers broadly to any distributed application that distinguishes between service providers (servers) and service requesters (clients).[26]
- Grid computing — "A form of distributed and parallel computing, whereby a 'super and virtual computer' is composed of a cluster of networked, loosely coupled computers acting in concert to perform very large tasks."
- Mainframe computer — Powerful computers used mainly by large organizations for critical applications, typically bulk data processing such as census, industry and consumer statistics, police and secret intelligence services, enterprise resource planning, and financial transaction processing.[27]
- Utility computing — The "packaging of computing resources, such as computation and storage, as a metered service similar to a traditional public utility, such as electricity."[28][29]
- Peer-to-peer means distributed architecture without the need for central coordination. Participants are both suppliers and consumers of resources (in contrast to the traditional client–server model).
- Cloud gaming—also known as on-demand gaming—is a way of delivering games to computers. Gaming data is stored in the provider's server, so that gaming is independent of client computers used to play the game.

Characteristics

Cloud computing exhibits the following key characteristics:

- **Agility** improves with users' ability to re-provision technological infrastructure resources.
- **Application programming interface** (API) accessibility to software that enables machines to interact with cloud software in the same way that a traditional user interface (e.g., a computer desktop) facilitates interaction between humans and computers. Cloud computing systems typically use Representational State Transfer (REST)-based APIs.
- **Cost** is claimed to be reduced, and in a public cloud delivery model capital expenditure is converted to operational expenditure.[30] This is purported to lower barriers to entry, as infrastructure is typically provided by a third-party and does not need to be purchased for one-time or infrequent intensive computing tasks. Pricing on a utility computing basis is fine-grained with usage-based options and fewer IT skills are required for implementation (in-house).[31] The e-FISCAL project's state of the art repository[32] contains several articles looking into cost aspects in more detail, most of them concluding that costs savings depend on the type of activities supported and the type of infrastructure available in-house.
- **Device and location independence**[33] enable users to access systems using a web browser regardless of their location or what device they are using (e.g., PC, mobile phone). As infrastructure is off-site (typically provided by a third-party) and accessed via the Internet, users can connect from anywhere.[31]
- **Virtualization** technology allows servers and storage devices to be shared and utilization be increased. Applications can be easily migrated from one physical server to another.
- **Multitenancy** enables sharing of resources and costs across a large pool of users thus allowing for:

    - **Centralization** of infrastructure in locations with lower costs (such as real estate, electricity, etc.)
    - **Peak-load capacity** increases (users need not engineer for highest possible load-levels)
    - **Utilisation and efficiency** improvements for systems that are often only 10–20% utilised.[11][34]
- **Reliability** is improved if multiple redundant sites are used, which makes well-designed cloud computing suitable for business continuity and disaster recovery.[35]
- **Scalability and elasticity** via dynamic ("on-demand") provisioning of resources on a fine-grained, self-service basis near real-time,[36][37] without users having to engineer for peak loads.[38][39][40]
- **Performance** is monitored, and consistent and loosely coupled architectures are constructed using web services as the system interface.[31]
- **Security** could improve due to centralization of data, increased security-focused resources, etc., but concerns can persist about loss of control over certain sensitive data, and the lack of security for stored kernels.[41] Security is often as good as or better than other traditional systems, in part because providers are able to devote resources to solving security issues that many customers cannot afford.[42] However, the complexity of security is greatly increased when data is distributed over a wider area or greater number of devices and in multi-tenant systems that are being shared by unrelated users. In addition, user access to security audit logs may be difficult or impossible. Private cloud installations are in part motivated by users' desire to retain control over the infrastructure and avoid losing control of information security.
- **Maintenance** of cloud computing applications is easier, because they do not need to be installed on each user's computer and can be accessed from different places.

The National Institute of Standards and Technology's definition of cloud computing identifies "five essential characteristics":

> *On-demand self-service.* A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
>
> *Broad network access.* Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
>
> *Resource pooling.* The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. ...
>
> *Rapid elasticity.* Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear unlimited and can be appropriated in any quantity at any time.
>
> *Measured service.* Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.
>
> —National Institute of Standards and Technology[2]

On-demand self-service

See also: Self-service provisioning for cloud computing services and Service catalogs for cloud computing services

On-demand self-service allows users to obtain, configure and deploy cloud services themselves using cloud service catalogues, without requiring the assistance of IT.[43][44] This feature is listed by the National Institute of Standards and Technology (NIST) as a characteristic of cloud computing.[2]

The self-service requirement of cloud computing prompts infrastructure vendors to create cloud computing templates, which are obtained from cloud service catalogues. Manufacturers of such templates or blueprints include BMC Software (BMC), with Service Blueprints as part of their cloud management platform[45] Hewlett-Packard (HP), which names its templates as HP Cloud Maps[46] RightScale[47] and Red Hat, which names its templates CloudForms.[48]

The templates contain predefined configurations used by consumers to set up cloud services. The templates or blueprints provide the technical information necessary to build ready-to-use clouds.[47] Each template includes specific configuration details for different cloud infrastructures, with information about servers for specific tasks such as hosting applications, databases, websites and so on.[47] The templates also include predefined Web service, the operating system, the database, security configurations and load balancing.[48]

Cloud computing consumers use cloud templates to move applications between clouds through a self-service portal. The predefined blueprints define all that an application requires to run in different environments. For example, a template could define how the same application could be deployed in cloud platforms based on Amazon Web Service, VMware or Red Hat.[49] The user organization benefits from cloud templates because the technical aspects of cloud configurations reside in the templates, letting users to deploy cloud services with a push of a button.[50][51] Cloud templates can also be used by developers to create a catalog of cloud services.[52]

Service models

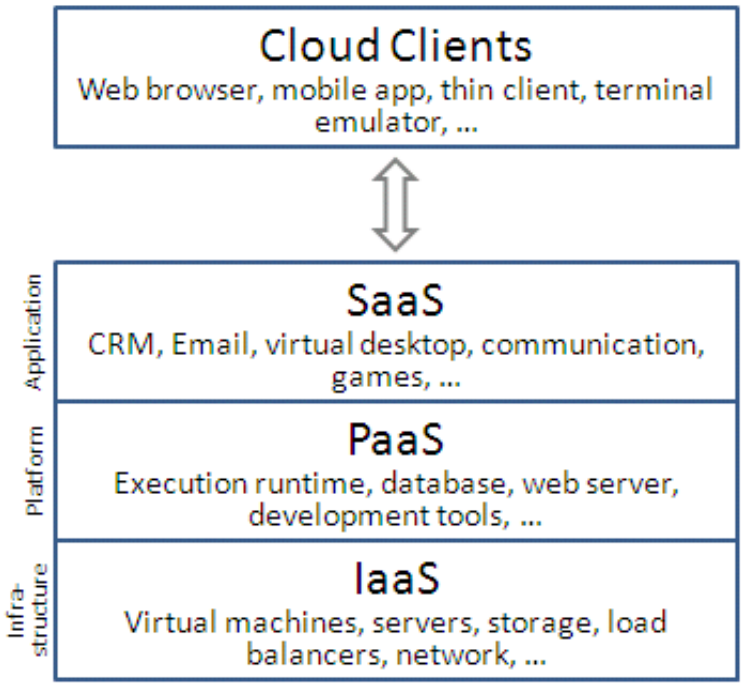
Cloud computing providers offer their services according to several fundamental models:[2][53] infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) where IaaS is the most basic and each higher model abstracts from the details of the lower models. Other key components in XaaS are described in a comprehensive taxonomy model published in 2009,[54] such as Strategy-as-a-Service, Collaboration-as-a-Service, Business Process-as-a-Service, Database-as-a-Service, etc. In 2012, network as a service (NaaS) and communication as a service (CaaS) were officially included by ITU (International Telecommunication Union) as part of the basic cloud computing models, recognized service categories of a telecommunication-centric cloud ecosystem.[55]

Infrastructure as a service (IaaS)

See also: Category:Cloud infrastructure

In the most basic cloud-service model, providers of IaaS offer computers - physical or (more often) virtual machines - and other resources. (A hypervisor, such as Xen or KVM, runs the virtual machines as guests. Pools of hypervisors within the cloud operational support-system can support large numbers of virtual machines and the ability to scale services up and down according to customers' varying requirements.) IaaS clouds often offer additional resources such as a virtual-machine disk image library, raw (block) and file-based storage, firewalls, load

balancers, IP addresses, virtual local area networks (VLANs), and software bundles.[56] IaaS-cloud providers supply these resources on-demand from their large pools installed in data centers. For wide-area connectivity, customers can use either the Internet or carrier clouds (dedicated virtual private networks).



To deploy their applications, cloud users install operating-system images and their application software on the cloud infrastructure. In this model, the cloud user patches and maintains the operating systems and the application software. Cloud providers typically bill IaaS services on a utility computing basis[*citation needed*]: cost reflects the amount of resources allocated and consumed.

Examples of IaaS providers include: Google Compute Engine, HP Cloud, Joyent, Linode, NaviSite, and ReadySpace Cloud Services.

The spending on cloud service is expected to show the largest increase in the IT marketplace, with North Africa and the Middle East having growth of over 20% through 2016, according to analysts at Gartner. The first cloud service in the United Arab Emirates for SMBs and enterprises was announced June 2013 when the leading telecom operator in the Middle East and Africa Etisalat launched its first cloud service in the UAE (http://www.dubaichronicle.com/2013/06/05/etisalat-launches-its-first-cloud-service-in-uae/). IaaS cloud model was believed to reduce IT costs up to 60% and time to market faster by up to 90%.

Cloud communications and cloud telephony, rather than replacing local computing infrastructure, replace local telecommunications infrastructure with Voice over IP and other off-site Internet services.

Platform as a service (PaaS)

Main article: Platform as a service

See also: Category:Cloud platforms

In the PaaS model, cloud providers deliver a computing platform, typically including operating system, programming language execution environment, database, and web server. Application developers can develop and run their software solutions on a cloud platform without the cost and complexity of buying and managing the underlying hardware and software layers. With some PaaS offers, the underlying computer and storage resources scale automatically to match application demand so that the cloud user does not have to allocate resources manually.

Examples of PaaS include: AWS Elastic Beanstalk, Cloud Foundry, Heroku, Force.com, EngineYard, Mendix, OpenShift, Google App Engine, AppScale, Windows Azure Cloud Services, OrangeScape and Jelastic.

Software as a service (SaaS)

Main article: Software as a service

In the business model using software as a service (SaaS), users are provided access to application software and databases. Cloud providers manage the infrastructure and platforms that run the applications. SaaS is sometimes referred to as "on-demand software" and is usually priced on a pay-per-use basis. SaaS providers generally price applications using a subscription fee.

In the SaaS model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. Cloud users do not manage the cloud infrastructure and platform where the application runs. This eliminates the need to install and run the application on the cloud user's own computers, which simplifies maintenance and support. Cloud applications are different from other applications in their scalability—which can be achieved by cloning tasks onto multiple virtual machines at run-time to meet changing work demand.[57] Load balancers distribute the work over the set of virtual machines. This process is transparent to the cloud user, who sees only a single access point. To accommodate a large number of cloud users, cloud applications can be *multitenant,* that is, any machine serves more than one cloud user organization. It is common to refer to special types of cloud based application software with a similar naming convention: desktop as a service, business process as a service, test environment as a service, communication as a service.

The pricing model for SaaS applications is typically a monthly or yearly flat fee per user,[58] so price is scalable and adjustable if users are added or removed at any point.[59]

Examples of SaaS include: Google Apps, Microsoft Office 365, Petrosoft, Onlive, GT Nexus, Marketo, Casengo, TradeCard, Salesforce and CallidusCloud.

Proponents claim SaaS allows a business the potential to reduce IT operational costs by outsourcing hardware and software maintenance and support to the cloud provider. This enables the business to reallocate IT operations costs away from hardware/software spending and personnel expenses, towards meeting other goals. In addition, with applications hosted centrally, updates can be released without the need for users to install new software. One drawback of SaaS is that the users' data are stored on the cloud provider's server. As a result, there could be unauthorized access to the data.

Network as a service (NaaS)

Main article: Network as a service

A category of cloud services where the capability provided to the cloud service user is to use network/transport connectivity services and/or inter-cloud network connectivity services.[60] NaaS involves the optimization of resource allocations by considering network and computing resources as a unified whole.[61]

Traditional NaaS services include flexible and extended VPN, and bandwidth on demand.[60] NaaS concept materialization also includes the provision of a virtual network service by the owners of the network infrastructure to a third party (VNP – VNO).[62][63]

Cloud clients

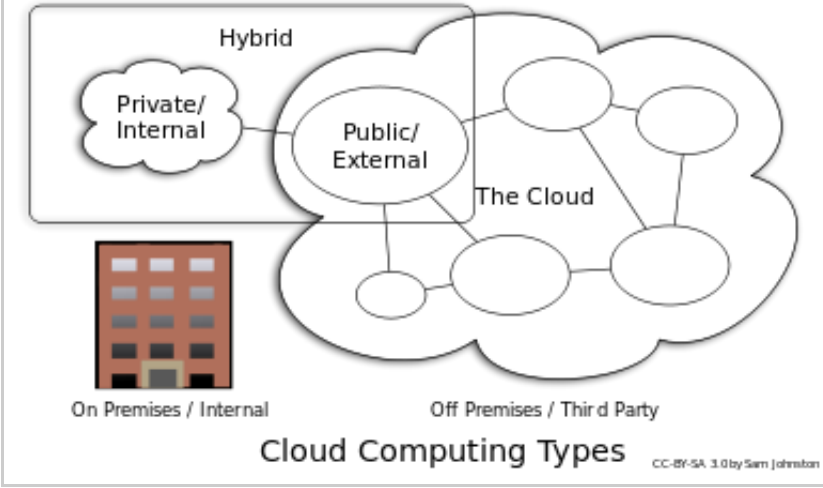
See also: Category:Cloud clients

Users access cloud computing using networked client devices, such as desktop computers, laptops, tablets and smartphones. Some of these devices - *cloud clients* - rely on cloud computing for all or a majority of their applications so as to be essentially useless without it. Examples are thin clients and the browser-based Chromebook. Many cloud applications do not require specific software on the client and instead use a web browser to interact with the cloud application. With Ajax and HTML5 these Web user interfaces can achieve a similar, or even better, look and feel to native applications. Some cloud applications, however, support specific client software dedicated to these applications (e.g., virtual desktop clients and most email clients). Some legacy applications (line of business applications that until now have been prevalent in thin client computing) are delivered via a screen-sharing technology.

Deployment models

Private cloud

Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third-party and hosted internally or externally.[2] Undertaking a private cloud project requires a significant level and degree of engagement to virtualize the business environment, and requires the organization to reevaluate decisions about existing resources. When done right, it can improve business, but every step in the project raises security issues that must be addressed to prevent serious vulnerabilities.[64]



Cloud computing types

They have attracted criticism because users "still have to buy, build, and manage them" and thus do not benefit from less hands-on management,[65] essentially "[lacking] the economic model that makes cloud computing such an intriguing concept".[66][67]

Comparison for SaaS

	Public cloud	Private cloud
Initial cost	Typically zero	Typically high
Running cost	Predictable	Unpredictable
Customization	Impossible	Possible
Privacy	No (Host has access to the data)	Yes
Single sign-on	Impossible	Possible
Scaling up	Easy while within defined limits	Laborious but no limits

Public cloud

A cloud is called a 'Public cloud' when the services are rendered over a network that is open for public use. Technically there is no difference between public and private cloud architecture, however, security consideration may be substantially different for services (applications, storage, and other resources) that are made available by a service provider for a public audience and when communication is effected over a non-trusted network. Generally, public cloud service providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access only via Internet (direct connectivity is not offered).[31]

Community cloud

Community cloud shares infrastructure between several organizations from a specific community with common concerns (security, compliance, jurisdiction, etc.), whether managed internally or by a third-party and hosted internally or externally. The costs are spread over fewer users than a public cloud (but more than a private cloud), so only some of the cost savings potential of cloud computing are realized.[2]

Hybrid cloud

Hybrid cloud is a composition of two or more clouds (private, community or public) that remain unique entities but are bound together, offering the benefits of multiple deployment models.[2] Such composition expands deployment options for cloud services, allowing IT organizations to use public cloud computing resources to meet temporary needs.[68] This capability enables hybrid clouds to employ cloud bursting for scaling across clouds.[2]

Cloud bursting is an application deployment model in which an application runs in a private cloud or data center and "bursts" to a public cloud when the demand for computing capacity increases. A primary advantage of cloud bursting and a hybrid cloud model is that an organization only pays for extra compute resources when they are needed.[69]

Cloud bursting enables data centers to create an in-house IT infrastructure that supports average workloads, and use cloud resources from public or private clouds, during spikes in processing demands.[70]

By utilizing "hybrid cloud" architecture, companies and individuals are able to obtain degrees of fault tolerance combined with locally immediate usability without dependency on internet connectivity. Hybrid cloud architecture requires both on-premises resources and off-site (remote) server-based cloud infrastructure.

Hybrid clouds lack the flexibility, security and certainty of in-house applications.[71] Hybrid cloud provides the flexibility of in house applications with the fault tolerance and scalability of cloud based services.

Architecture

Cloud architecture,[72] the systems architecture of the software systems involved in the delivery of cloud computing, typically involves multiple *cloud components* communicating with each other over a loose coupling mechanism such as a messaging queue. Elastic provision implies intelligence in the use of tight or loose coupling as applied to mechanisms such as these and others.

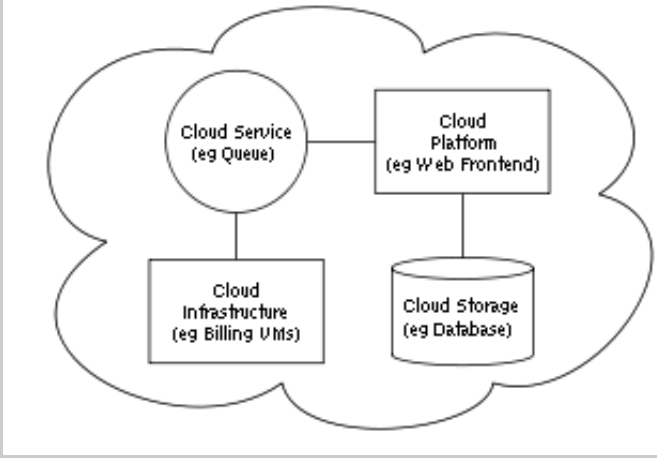
The Intercloud

Main article: Intercloud

The Intercloud[73] is an interconnected global "cloud of clouds"[74][75] and an extension of the Internet "network of networks" on which it is based.[76][77][78]

Cloud engineering

Cloud engineering is the application of engineering disciplines to cloud computing. It brings a systematic approach to the high-level concerns of commercialisation, standardisation, and governance in conceiving, developing, operating and maintaining cloud computing systems. It is a multidisciplinary method encompassing contributions from diverse areas such as systems, software, web, performance, information, security, platform, risk, and quality engineering.



Cloud computing sample architecture

Issues

Threats and opportunities of the cloud

Critical voices including GNU project initiator Richard Stallman and Oracle founder Larry Ellison warned that the whole concept is rife with privacy and ownership concerns and constitute merely a fad.[79]

However, cloud computing continues to gain steam[80] with 56% of the major European technology decision-makers estimate that the cloud is a priority in 2013 and 2014, and the cloud budget may reach 30% of the overall IT budget.[*citation needed*][81]

According to the *TechInsights Report 2013: Cloud Succeeds* based on a survey, the cloud implementations generally meets or exceedes expectations across major service models, such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS)".[82]

Several deterrents to the widespread adoption of cloud computing remain. Among them, are: reliability, availability of services and data, security, complexity, costs, regulations and legal issues, performance, migration, reversion, the lack of standards, limited customization and issues of privacy. The *cloud* offers many strong points: infrastructure flexibility, faster deployment of applications and data, cost control, adaptation of cloud resources to real needs, improved productivity, etc. The early 2010s cloud market is dominated by software and services in SaaS mode and IaaS (infrastructure), especially the private cloud. PaaS and the public cloud are further back.

Privacy

Privacy advocates have criticized the cloud model for giving hosting companies' greater ease to control—and thus, to monitor at will—communication between host company and end user, and access user data (with or without permission). Instances such as the secret NSA program, working with AT&T, and Verizon, which recorded over 10 million telephone calls between American citizens, causes uncertainty among privacy advocates, and the greater powers it gives to telecommunication companies to monitor user activity.[83][84] A cloud service provider (CSP) can complicate data privacy because of the extent of virtualization (virtual machines) and cloud storage used to implement cloud service.[85] CSP operations, customer or tenant data may not remain on the same system, or in the

same data center or even within the same provider's cloud; this can lead to legal concerns over jurisdiction. While there have been efforts (such as US-EU Safe Harbor) to "harmonise" the legal environment, providers such as Amazon still cater to major markets (typically the United States and the European Union) by deploying local infrastructure and allowing customers to select "availability zones."[86] Cloud computing poses privacy concerns because the service provider can access the data that is on the cloud at any time. It could accidentally or deliberately alter or even delete information.[87]

Compliance

To comply with regulations including FISMA, HIPAA, and SOX in the United States, the Data Protection Directive in the EU and the credit card industry's PCI DSS, users may have to adopt *community* or *hybrid* deployment modes that are typically more expensive and may offer restricted benefits. This is how Google is able to "manage and meet additional government policy requirements beyond FISMA"[88][89] and Rackspace Cloud or QubeSpace are able to claim PCI compliance.[90]

Many providers also obtain a SAS 70 Type II audit, but this has been criticised on the grounds that the hand-picked set of goals and standards determined by the auditor and the auditee are often not disclosed and can vary widely.[91] Providers typically make this information available on request, under non-disclosure agreement.[92][93]

Customers in the EU contracting with cloud providers outside the EU/EEA have to adhere to the EU regulations on export of personal data.[94]

U.S. Federal Agencies have been directed by the Office of Management and Budget to use a process called FedRAMP (Federal Risk and Authorization Management Program) to assess and authorize cloud products and services. Federal CIO Steven VanRoekel issued a memorandum to federal agency Chief Information Officers on December 8, 2011 defining how federal agencies should use FedRAMP. FedRAMP consists of a subset of NIST Special Publication 800-53 security controls specifically selected to provide protection in cloud environments. A subset has been defined for the FIPS 199 low categorization and the FIPS 199 moderate categorization. The FedRAMP program has also established a Joint Accreditation Board (JAB) consisting of Chief Information Officers from DoD, DHS and GSA. The JAB is responsible for establishing accreditation standards for 3rd party organizations who perform the assessments of cloud solutions. The JAB also reviews authorization packages, and may grant provisional authorization (to operate). The federal agency consuming the service still has final responsibility for final authority to operate.[95]

A multitude of laws and regulations have forced specific compliance requirements onto many companies that collect, generate or store data. These policies may dictate a wide array of data storage policies, such as how long information must be retained, the process used for deleting data, and even certain recovery plans. Below are some examples of compliance laws or regulations.

- In the United States, the Health Insurance Portability and Accountability Act (HIPAA) requires a contingency plan that includes, data backups, data recovery, and data access during emergencies.
- The privacy laws of the Switzerland demand that private data, including emails, be physically stored in the Switzerland.
- In the United Kingdom, the Civil Contingencies Act of 2004 sets forth guidance for a Business contingency plan that includes policies for data storage.

In a virtualized cloud computing environment, customers may never know exactly where their data is stored. In fact, data may be stored across multiple data centers in an effort to improve reliability, increase performance, and provide redundancies. This geographic dispersion may make it more difficult to ascertain legal jurisdiction if disputes arise.[96]

Legal

As with other changes in the landscape of computing, certain legal issues arise with cloud computing, including trademark infringement, security concerns and sharing of proprietary data resources.

The Electronic Frontier Foundation has criticized the United States government for considering during the Megaupload seizure process that people lose property rights by storing data on a cloud computing service.[97]

One important but not often mentioned problem with cloud computing is the problem of who is in "possession" of the data. If a cloud company is the possessor of the data, the possessor has certain legal rights. If the cloud company is the "custodian" of the data, then a different set of rights would apply. The next problem in the legalities of cloud computing is the problem of legal ownership of the data. Many Terms of Service agreements are silent on the question of ownership.[98]

These legal issues are not confined to the time period in which the cloud based application is actively being used. There must also be consideration for what happens when the provider-customer relationship ends. In most cases, this event will be addressed before an application is deployed to the cloud. However, in the case of provider insolvencies or bankruptcy the state of the data may become blurred.[96]

Vendor lock-in

Because cloud computing is still relatively new, standards are still being developed.[99] Many cloud platforms and services are proprietary, meaning that they are built on the specific standards, tools and protocols developed by a particular vendor for its particular cloud offering.[99] This can make migrating off a proprietary cloud platform prohibitively complicated and expensive.[99]

Three types of vendor lock-in can occur with cloud computing:[100]

- Platform lock-in: cloud services tend to be built on one of several possible virtualization platforms, for example VMWare or Xen. Migrating from a cloud provider using one platform to a cloud provider using a different platform could be very complicated.
- Data lock-in: since the cloud is still new, standards of ownership, i.e. who actually owns the data once it lives on a cloud platform, are not yet developed, which could make it complicated if cloud computing users ever decide to move data off of a cloud vendor's platform.
- Tools lock-in: if tools built to manage a cloud environment are not compatible with different kinds of both virtual and physical infrastructure, those tools will only be able to manage data or apps that live in the vendor's particular cloud environment.

Heterogeneous cloud computing is described as a type of cloud environment that prevents vendor lock-in, and aligns with enterprise data centers that are operating hybrid cloud models.[101] The absence of vendor lock-in lets cloud administrators select his or her choice of hypervisors for specific tasks, or to deploy virtualized infrastructures to other enterprises without the need to consider the flavor of hypervisor in the other enterprise.[102]

A heterogeneous cloud is considered one that includes on-premise private clouds, public clouds and software-as-a-service clouds. Heterogeneous clouds can work with environments that are not virtualized, such as traditional data centers.[103] Heterogeneous clouds also allow for the use of piece parts, such as hypervisors, servers, and storage, from multiple vendors.[104]

Cloud piece parts, such as cloud storage systems, offer APIs but they are often incompatible with each other.[105] The result is complicated migration between backends, and makes it difficult to integrate data spread across various locations.[105] This has been described as a problem of vendor lock-in.[105] The solution to this is for clouds to adopt common standards.[105]

Heterogeneous cloud computing differs from homogeneous clouds, which have been described as those using consistent building blocks supplied by a single vendor.[106] Intel General Manager of high-density computing, Jason Waxman, is quoted as saying that a homogenous system of 15,000 servers would cost $6 million more in capital expenditure and use 1 megawatt of power.[106]

Open source

See also: Category:Free software for cloud computing

Open-source software has provided the foundation for many cloud computing implementations, prominent examples being the Hadoop framework[107] and VMware's Cloud Foundry.[108] In November 2007, the Free Software Foundation released the Affero General Public License, a version of GPLv3 intended to close a perceived legal loophole associated with free software designed to run over a network.[109]

Open standards

See also: Category:Cloud standards

Most cloud providers expose APIs that are typically well-documented (often under a Creative Commons license[110]) but also unique to their implementation and thus not interoperable. Some vendors have adopted others' APIs and there are a number of open standards under development, with a view to delivering interoperability and portability.[111] As of November 2012, the Open Standard with broadest industry support is probably OpenStack, founded in 2010 by NASA and Rackspace, and now governed by the OpenStack Foundation.[112] OpenStack supporters include AMD, Intel, Canonical, SUSE Linux, Red Hat, Cisco, Dell, HP, IBM, Yahoo and now VMware.[113]

Security

Main article: Cloud computing security

As cloud computing is achieving increased popularity, concerns are being voiced about the security issues introduced through adoption of this new model.[1] The effectiveness and efficiency of traditional protection mechanisms are being reconsidered as the characteristics of this innovative deployment model can differ widely from those of traditional architectures.[114] An alternative perspective on the topic of cloud security is that this is but another, although quite broad, case of "applied security" and that similar security principles that apply in shared multi-user mainframe security models apply with cloud security.[115]

The relative security of cloud computing services is a contentious issue that may be delaying its adoption.[116] Physical control of the Private Cloud equipment is more secure than having the equipment off site and under someone else's control. Physical control and the ability to visually inspect data links and access ports is required in order to ensure data links are not compromised. Issues barring the adoption of cloud computing are due in large part to the private and public sectors' unease surrounding the external management of security-based services. It is the very nature of cloud computing-based services, private or public, that promote external management of provided services. This delivers great incentive to cloud computing service providers to prioritize building and maintaining strong management of secure services.[117] Security issues have been categorised into sensitive data access, data segregation, privacy, bug exploitation, recovery, accountability, malicious insiders, management console security, account control, and multi-tenancy issues. Solutions to various cloud security issues vary, from cryptography, particularly public key infrastructure (PKI), to use of multiple cloud providers, standardisation of APIs, and improving virtual machine support and legal support.[114][118][119]

Cloud computing offers many benefits, but is vulnerable to threats. As cloud computing uses increase, it is likely that more criminals find new ways to exploit system vulnerabilities. Many underlying challenges and risks in cloud computing increase the threat of data compromise. To mitigate the threat, cloud computing stakeholders should invest heavily in risk assessment to ensure that the system encrypts to protect data, establishes trusted foundation to secure the platform and infrastructure, and builds higher assurance into auditing to strengthen compliance. Security concerns must be addressed to maintain trust in cloud computing technology.[1]

Sustainability

Although cloud computing is often assumed to be a form of *green computing*, no published study substantiates this assumption.[120] Citing the servers' effects on the environmental effects of cloud computing, in areas where climate favors natural cooling and renewable electricity is readily available, the environmental effects will be more moderate. (The same holds true for "traditional" data centers.) Thus countries with favorable conditions, such as Finland,[121] Sweden and Switzerland,[122] are trying to attract cloud computing data centers. Energy efficiency in cloud computing can result from energy-aware scheduling and server consolidation.[123] However, in the case of distributed clouds over data centers with different source of energies including renewable source of energies, a small compromise on energy consumption reduction could result in high carbon footprint reduction.[124]

Abuse

As with privately purchased hardware, customers can purchase the services of cloud computing for nefarious purposes. This includes password cracking and launching attacks using the purchased services.[125] In 2009, a banking trojan illegally used the popular Amazon service as a command and control channel that issued software updates and malicious instructions to PCs that were infected by the malware.[126]

IT governance

Main article: Corporate governance of information technology

The introduction of cloud computing requires an appropriate IT governance model to ensure a secured computing environment and to comply with all relevant organizational information technology policies.[127][128] As such, organizations need a set of capabilities that are essential when effectively implementing and managing cloud services, including demand management, relationship management, data security management, application lifecycle

management, risk and compliance management.[129] A danger lies with the explosion of companies joining the growth in cloud computing by becoming providers. However, many of the infrastructural and logistical concerns regarding the operation of cloud computing businesses are still unknown. This over-saturation may have ramifications for the industry as whole.[130]

Consumer end storage

The increased use of cloud computing could lead to a reduction in demand for high storage capacity consumer end devices, due to cheaper low storage devices that stream all content via the cloud becoming more popular.[*citation needed*] In a Wired article, Jake Gardner explains that while unregulated usage is beneficial for IT and tech moguls like Amazon, the anonymous nature of the cost of consumption of cloud usage makes it difficult for business to evaluate and incorporate it into their business plans.[130] The popularity of cloud and cloud computing in general is so quickly increasing among all sorts of companies, that in May 2013, through its company Amazon Web Services, Amazon started a certification program for cloud computing professionals. (http://www.dubaichronicle.com/2013/05/13/amazon-starts-a-certification-program-for-cloud-computing-pros/)

Ambiguity of terminology

Outside of the information technology and software industry, the term "cloud" can be found to reference a wide range of services, some of which fall under the category of cloud computing, while others do not. The cloud is often used to refer to a product or service that is discovered, accessed and paid for over the Internet, but is not necessarily a computing resource. Examples of service that are sometimes referred to as "the cloud" include, but are not limited to, crowd sourcing, cloud printing, crowd funding, cloud manufacturing.[131][132]

Performance interference and noisy neighbors

Due to its multi-tenant nature and resource sharing, Cloud computing must also deal with the "noisy neighbor" effect. This effect in essence indicates that in a shared infrastructure, the activity of a virtual machine on a neighboring core on the same physical host may lead to increased performance degradation of the VMs in the same physical host, due to issues such as e.g. cache contamination. Due to the fact that the neighboring VMs may be activated or deactivated at arbitrary times, the result is an increased variation in the actual performance of Cloud resources. This effect seems to be dependent also on the nature of the applications that run inside the VMs but also other factors such as scheduling parameters and the careful selection may lead to optimized assignment in order to minimize the phenomenon. This has also led to difficulties in comparing various cloud providers on cost and performance using traditional benchmarks for service and application performance, as the time period and location in which the benchmark is performed can result in widely varied results.[133]

Research

Many universities, vendors, Institutes and government organizations are investing in research around the topic of cloud computing:[134][135]

- In October 2007, the Academic Cloud Computing Initiative (ACCI) was announced as a multi-university project designed to enhance students' technical knowledge to address the challenges of cloud computing.[136]

- In April 2009, UC Santa Barbara released the first open source platform-as-a-service, AppScale, which is capable of running Google App Engine applications at scale on a multitude of infrastructures.
- In April 2009, the St Andrews Cloud Computing Co-laboratory was launched, focusing on research in the important new area of cloud computing. Unique in the UK, StACC aims to become an international centre of excellence for research and teaching in cloud computing and provides advice and information to businesses interested in cloud-based services.[137]
- In October 2010, the TClouds (Trustworthy Clouds) project was started, funded by the European Commission's 7th Framework Programme. The project's goal is to research and inspect the legal foundation and architectural design to build a resilient and trustworthy cloud-of-cloud infrastructure on top of that. The project also develops a prototype to demonstrate its results.[138]
- In December 2010, the TrustCloud research project [139][140] was started by HP Labs Singapore to address transparency and accountability of cloud computing via detective, data-centric approaches[141] encapsulated in a five-layer TrustCloud Framework. The team identified the need for monitoring data life cycles and transfers in the cloud,[139] leading to the tackling of key cloud computing security issues such as cloud data leakages, cloud accountability and cross-national data transfers in transnational clouds.
- In June 2011, two Indian Universities i.e. University of Petroleum and Energy Studies and University of Technology and Management introduced cloud computing as a subject in India, in collaboration with IBM.[142]
- In July 2011, the High Performance Computing Cloud (HPCCLoud) project was kicked-off aiming at finding out the possibilities of enhancing performance on cloud environments while running the scientific applications - development of HPCCLoud Performance Analysis Toolkit which was funded by CIM-Returning Experts Programme - under the coordination of Prof. Dr. Shajulin Benedict.
- In June 2011, the Telecommunications Industry Association developed a Cloud Computing White Paper, to analyze the integration challenges and opportunities between cloud services and traditional U.S. telecommunications standards.[143]
- In December 2012, a study released by Microsoft and the International Data Corporation (IDC)showed that millions of cloud-skilled workers would be needed (http://www.dubaichronicle.com/2012/12/22/cloud-skilled-it-workers/). Millions of cloud-related IT jobs are sitting open and millions more will open in the coming couple of years, due to a shortage in cloud-certified IT workers.
- In February 2013, the BonFIRE project launched a multi-site cloud experimentation and testing facility. The facility provides transparent access to cloud resources, with the control and observability necessary to engineer future cloud technologies, in a way that is not restricted, for example, by current business models.[144]
- In April 2013, A 2013 report by IT research and advisory firm Gartner., Inc. says that app developers will embrace cloud services (http://www.dubaichronicle.com/2013/05/01/mobile-app-development-cloud-services/), predicting that in three years, 40% of the mobile app development projects will use cloud backed services. Cloud mobile backed services offer a new kind of PaaS, used to enable the development of mobile

apps.

Early references in popular culture

In the 1966 *Star Trek* episode "Miri," Dr. McCoy, while stationed planetside, uses the computer of the orbiting Enterprise to process the data gathered by his portable equipment.

See also

- Cloud collaboration
- Cloud computing comparison
- Cloud telephony
- List of cloud computing conferences
- Mobile cloud computing
- Web operating system

References

1. ^ ***a b c*** Mariana Carroll, Paula Kotzé, Alta van der Merwe. 2012. Securing Virtual and Cloud Environments. In: Cloud Computing and Services Science, Service Science: Research and Innovations in the Service Economy), edited by I. Ivanov et al., DOI 10.1007/978-1-4614-2326-3 4, © Springer Science+Business Media, LLC 2012
2. ^ ***a b c d e f g h*** "The NIST Definition of Cloud Computing" (http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf). National Institute of Standards and Technology. Retrieved 24 July 2011.
3. ^ ***a b*** "What is Cloud Computing?" (http://aws.amazon.com/what-is-cloud-computing/). *Amazon Web Services*. 2013-3-19. Retrieved 2013-3-20.
4. ^ "Baburajan, Rajani, "The Rising Cloud Storage Market Opportunity Strengthens Vendors," infoTECH, August 24, 2011" (http://it.tmcnet.com/channels/cloud-storage/articles/211183-rising-cloud-storage-market-opportunity-strengthens-vendors.htm). It.tmcnet.com. 2011-08-24. Retrieved 2011-12-02.
5. ^ Oestreich, Ken, (2010-11-15). "Converged Infrastructure" (http://www.thectoforum.com/content/converged-infrastructure-0). *CTO Forum*. Thectoforum.com. Retrieved 2011-12-02.
6. ^ Strachey, Christopher (June 1959). "Time Sharing in Large Fast Computers". *Proceedings of the International Conference on Information processing, UNESCO*. paper B.2.19: 336–341.
7. ^ Simson Garfinkel (3 October 2011). "The Cloud Imperative" (http://www.technologyreview.com/news/425623/the-cloud-imperative/). *Technology Review* (MIT). Retrieved 31 May 2013.
8. ^ Ryan; Falvey; Merchant (October 2011). "Regulation of the Cloud in India" (http://ssrn.com/abstract=1941494). *Journal of Internet Law* **15** (4)
9. ^ "July, 1993 meeting report from the IP over ATM working group of the IETF" (http://mirror.switch.ch/ftp/doc/ietf/ipatm/atm-minutes-93jul.txt). CH: Switch. Retrieved 2010-08-22.
10. ^ Corbató, Fernando J. "An Experimental Time-Sharing System" (http://larch-www.lcs.mit.edu:8001/~corbato/sjcc62/). *SJCC Proceedings*. MIT. Retrieved 3 July 2012.
11. ^ ***a b*** "Jeff Bezos' Risky Bet" (http://www.businessweek.com/magazine/content/06_46/b4009001.htm). *Business Week*
12. ^ "Amazon's early efforts at cloud computing partly accidental" (http://itknowledgeexchange.techtarget.com/cloud-computing/2010/06/17/amazons-early-efforts-at-cloud-computing-partly-accidental/). *IT Knowledge Exchange*. Tech Target. 2010-06-17
13. ^ B Rochwerger, J Caceres, RS Montero, D Breitgand, E Elmroth, A Galis, E Levy, IM Llorente, K Nagin, Y Wolfsthal. E Elmroth. J Caceres. M Ben-Yehuda. W Emmerich. F Galan. "The RESERVOIR Model and

Architecture for Open Federated Cloud Computing", IBM Journal of Research and Development, Vol. 53, No. 4. (2009)
14. ^ D Kyriazis, A Menychtas, G Kousiouris, K Oberle, T Voith, M Boniface, E Oliveros, T Cucinotta, S Berger, "A Real-time Service Oriented Infrastructure", International Conference on Real-Time and Embedded Systems (RTES 2010), Singapore, November 2010
15. ^ Keep an eye on cloud computing (http://www.networkworld.com/newsletters/itlead/2008/070708itlead1.html), Amy Schurr, Network World, 2008-07-08, citing the Gartner report, "Cloud Computing Confusion Leads to Opportunity". Retrieved 2009-09-11.
16. ^ Gartner Says Worldwide IT Spending On Pace to Surpass Trillion in 2008 (http://www.gartner.com/it/page.jsp?id=742913), Gartner, 2008-08-18. Retrieved 2009-09-11.
17. ^ "Launch of IBM Smarter Computing" (https://www-304.ibm.com/connections/blogs/IBMSmarterSystems/date/201102?lang=en_us). Retrieved 1 March 2011.
18. ^ Andreas Tolk. 2006. What Comes After the Semantic Web - PADS Implications for the Dynamic Web. 20th Workshop on Principles of Advanced and Distributed Simulation (PADS '06). IEEE Computer Society, Washington, DC, USA
19. ^ "Cloud Computing: Clash of the clouds" (http://www.economist.com/displaystory.cfm?story_id=14637206). The Economist. 2009-10-15. Retrieved 2009-11-03.
20. ^ "Gartner Says Cloud Computing Will Be As Influential As E-business" (http://www.gartner.com/it/page.jsp?id=707508). Gartner. Retrieved 2010-08-22.
21. ^ Gruman, Galen (2008-04-07). "What cloud computing really means" (http://www.infoworld.com/d/cloud-computing/what-cloud-computing-really-means-031). *InfoWorld*. Retrieved 2009-06-02.
22. ^ "The economy is flat so why are financials Cloud vendors growing at more than 90 percent per annum?" (http://www.fsn.co.uk/channel_outsourcing/the_economy_is_flat_so_why_are_financials_cloud_vendors_growing_at_more_than_90_percent_per_annum#.UbmtsPlJPGA/). FSN. March 5, 2013.
23. ^ Figure 8, "A network 70 is shown schematically as a cloud", US Patent 5,485,455, column 17, line 22, filed Jan 28, 1994
24. ^ Figure 1, "the cloud indicated at 49 in Fig. 1.", US Patent 5,790,548, column 5 line 56-57, filed April 18, 1996
25. ^ ***a b c*** HAMDAQA, Mohammad (2012). *Cloud Computing Uncovered: A Research Landscape* (http://www.stargroup.uwaterloo.ca/~mhamdaqa/publications/Cloud_Computing_Uncovered.pdf). Elsevier Press. pp. 41–85. ISBN 0-12-396535-7.
26. ^ "Distributed Application Architecture" (http://java.sun.com/developer/Books/jdbc/ch07.pdf). Sun Microsystem. Retrieved 2009-06-16.
27. ^ "Sun CTO: Cloud computing is like the mainframe" (http://itknowledgeexchange.techtarget.com/mainframe-blog/sun-cto-cloud-computing-is-like-the-mainframe/). Itknowledgeexchange.techtarget.com. 2009-03-11. Retrieved 2010-08-22.
28. ^ "It's probable that you've misunderstood 'Cloud Computing' until now" (http://portal.acm.org/citation.cfm?id=1496091.1496100&coll=&dl=ACM&CFID=21518680&CFTOKEN=18800807). TechPluto. Retrieved 2010-09-14.
29. ^ Danielson, Krissi (2008-03-26). "Distinguishing Cloud Computing from Utility Computing" (http://www.ebizq.net/blogs/saasweek/2008/03/distinguishing_cloud_computing/). Ebizq.net. Retrieved 2010-08-22.
30. ^ "Recession Is Good For Cloud Computing – Microsoft Agrees" (http://www.cloudave.com/link/recession-is-good-for-cloud-computing-microsoft-agrees). CloudAve. Retrieved 2010-08-22.
31. ^ ***a b c d*** "Defining "Cloud Services" and "Cloud Computing"" (http://blogs.idc.com/ie/?p=190). IDC. 2008-09-23. Retrieved 2010-08-22.
32. ^ "e-FISCAL project state of the art repository" (http://www.efiscal.eu/state-of-the-art).
33. ^ Farber, Dan (2008-06-25). "The new geek chic: Data centers" (http://news.cnet.com/8301-13953_3-9977049-80.html). CNET News. Retrieved 2010-08-22.
34. ^ He, Sijin; L. Guo, Y. Guo, M. Ghanem,. *Improving Resource Utilisation in the Cloud Environment Using Multivariate Probabilistic Models* (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6253553). 2012 2012 IEEE 5th International Conference on Cloud Computing (CLOUD). pp. 574–581. doi:10.1109/CLOUD.2012.66 (http://dx.doi.org/10.1109%2FCLOUD.2012.66). ISBN 978-1-4673-2892-0.

35. ^ King, Rachael (2008-08-04). "Cloud Computing: Small Companies Take Flight" (http://www.businessweek.com/technology/content/aug2008/tc2008083_619516.htm). Businessweek. Retrieved 2010-08-22.
36. ^ Mao, Ming; M. Humphrey (2012). "A Performance Study on the VM Startup Time in the Cloud" (http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6253534&isnumber=6253471). *Proceedings of 2012 IEEE 5th International Conference on Cloud Computing (Cloud2012)*: 423. doi:10.1109/CLOUD.2012.103 (http://dx.doi.org/10.1109%2FCLOUD.2012.103). ISBN 978-1-4673-2892-0.
37. ^ He, Sijin; L. Guo, Y. Guo (2011). "Real Time Elastic Cloud Management for Limited Resources" (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6008763). *Proceedings of 2011 IEEE 4th International Conference on Cloud Computing (Cloud2011)*: 622–629. doi:10.1109/CLOUD.2011.47 (http://dx.doi.org/10.1109%2FCLOUD.2011.47). ISBN 978-0-7695-4460-1.
38. ^ "Defining and Measuring Cloud Elasticity" (http://digbib.ubka.uni-karlsruhe.de/volltexte/1000023476). KIT Software Quality Departement. Retrieved 13 August 2011.
39. ^ "Economies of Cloud Scale Infrastructure" (http://www.youtube.com/watch?v=nfDsY3f4nVI). Cloud Slam 2011. Retrieved 13 May 2011.
40. ^ He, Sijin; L. Guo, Y. Guo, C. Wu, M. Ghanem, R. Han. *Elastic Application Container: A Lightweight Approach for Cloud Resource Provisioning* (http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6184989). 2012 IEEE 26th International Conference on Advanced Information Networking and Applications (AINA). pp. 15–22. doi:10.1109/AINA.2012.74 (http://dx.doi.org/10.1109%2FAINA.2012.74). ISBN 978-1-4673-0714-7.
41. ^ "Encrypted Storage and Key Management for the cloud" (http://www.cryptoclarity.com/CryptoClarityLLC/Welcome/Entries/2009/7/23_Encrypted_Storage_and_Key_Management_for_the_cloud.html). Cryptoclarity.com. 2009-07-30. Retrieved 2010-08-22.
42. ^ Mills, Elinor (2009-01-27). "Cloud computing security forecast: Clear skies" (http://news.cnet.com/8301-1009_3-10150569-83.html). CNET News. Retrieved 2010-08-22.
43. ^ David Perera (2012-07-12). "The real obstacle to federal cloud computing" (http://www.fiercegovernmentit.com/story/real-obstacle-federal-cloud-computing/2012-07-12). FierceGovernmentIT. Retrieved 2012-12-15.
44. ^ "Top 10 Reasons why Startups should Consider Cloud" (http://cloudstory.in/2012/07/top-10-reasons-why-startups-should-consider-cloud/). Cloudstory.in. 2012-09-05. Retrieved 2012-12-15.
45. ^ "BMC Service Catalog Enforces Workload Location" (http://www.eweek.com/reviews/bmc-service-catalog-enforces-workload-location). eweek.com. 2011-08-02. Retrieved 2013-03-10.
46. ^ "HP's Turn-Key Private Cloud - Application Development Trends" (http://adtmag.com/blogs/watersworks/2010/08/hps-turn-key-private-cloud.aspx). Adtmag.com. 2010-08-30. Retrieved 2012-12-15.
47. ^ ***a*** ***b*** ***c*** Babcock, Charles (2011-06-03). "RightScale Launches App Store For Infrastructure - Cloud-computing" (http://www.informationweek.com/news/cloud-computing/infrastructure/229900165). Informationweek.com. Retrieved 2012-12-15.
48. ^ ***a*** ***b*** "Red Hat launches hybrid cloud management software - Open Source" (http://www.techworld.com.au/article/426861/red_hat_launches_hybrid_cloud_management_software). Techworld. 2012-06-06. Retrieved 2012-12-15.
49. ^ Brown, Rodney (April 10, 2012). "Spinning up the instant cloud" (http://www.cloudecosystem.com/author.asp?section_id=1873&doc_id=242031). *CloudEcosystem*.
50. ^ Riglian, Adam (December 1, 2011). "VIP Art Fair picks OpDemand over RightScale for IaaS management" (http://searchcloudapplications.techtarget.com/news/2240111861/VIP-Art-Fair-picks-OpDemand-over-RightScale-for-IaaS-management). *Search Cloud Applications*. TechTarget. Retrieved January 25, 2013.
51. ^ Samson, Ted (April 10, 2012). "HP advances public cloud as part of ambitious hybrid cloud strategy" (http://www.infoworld.com/t/hybrid-cloud/hp-advances-public-cloud-part-of-ambitious-hybrid-cloud-strategy-190524). InfoWorld. Retrieved 2012-12-14.
52. ^ "HP Cloud Maps can ease application automation" (http://www.siliconindia.com/shownews/HP_Cloud_Maps_can_Ease_Application_Automation-nid-103866-cid-7.html). *SiliconIndia*. Retrieved 22 January 2013.
53. ^ Voorsluys, William; Broberg, James; Buyya, Rajkumar (February 2011). "Introduction to Cloud Computing" (http://media.johnwiley.com.au/product_data/excerpt/90/04708879/0470887990-180.pdf). In R. Buyya, J. Broberg,

A.Goscinski. *Cloud Computing: Principles and Paradigms*. New York, USA: Wiley Press. pp. 1–44. ISBN 978-0-470-88799-8.
54. ^ "Tony Shan, "Cloud Taxonomy and Ontology"" (http://cloudonomic.blogspot.com/2009/02/cloud-taxonomy-and-ontology.html). February 2009. Retrieved 2 February 2009.
55. ^ "ITU-T NEWSLOG - CLOUD COMPUTING AND STANDARDIZATION: TECHNICAL REPORTS PUBLISHED" (http://www.itu.int/ITU-T/newslog/Cloud+Computing+And+Standardization+Technical+Reports+Published.aspx). International Telecommunication Union (ITU). Retrieved 16 December 2012.
56. ^ Amies, Alex; Sluiman, Harm; Tong, Qiang Guo; Liu, Guo Ning (July 2012). "Infrastructure as a Service Cloud Concepts" (http://www.ibmpressbooks.com/bookstore/product.asp?isbn=9780133066845). *Developing and Hosting Applications on the Cloud*. IBM Press. ISBN 978-0-13-306684-5.
57. ^ Hamdaqa, Mohammad. *A Reference Model for Developing Cloud Applications* (http://www.stargroup.uwaterloo.ca/~mhamdaqa/publications/A%20REFERENCEMODELFORDEVELOPINGCLOUD%20APPLICATIONS.pdf).
58. ^ Chou, Timothy. *Introduction to Cloud Computing: Business & Technology* (http://www.scribd.com/doc/64699897/Introduction-to-Cloud-Computing-Business-and-Technology).
59. ^ "HVD: the cloud's silver lining" (http://www.intrinsictechnology.co.uk/FileUploads/HVD_Whitepaper.pdf). Intrinsic Technology. Retrieved 30 August 2012.
60. ^ ***a*** ***b*** "ITU Focus Group on Cloud Computing - Part 1" (http://www.itu.int/en/ITU-T/focusgroups/cloud/Documents/FG-coud-technical-report.zip). International Telecommunication Union (ITU) TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU. Retrieved 16 December 2012.
61. ^ "Cloud computing in Telecommunications" (http://www.ericsson.com/res/thecompany/docs/publications/ericsson_review/2010/cloudcomputing.pdf). Ericsson. Retrieved 16 December 2012.
62. ^ "Network Virtualisation – Opportunities and Challenges" (http://archive.eurescom.eu/~pub/deliverables/documents/P1900-series/P1956/D1/P1956-D1.pdf). Eurescom. Retrieved 16 December 2012.
63. ^ "The role of virtualisation in future network architectures" (http://www.change-project.eu/fileadmin/publications/Presentations/CHANGE_-_The_role_of_virtualisation_in_future_network_infrastructures_-_Warsaw_cluster_workshop_contribution.pdf). Change Project. Retrieved 16 December 2012.
64. ^ "Is a Private Cloud Really More Secure?" (http://www.dell.com/Learn/us/en/rc956904/large-business/private-cloud-more-secure). Dell.com. Retrieved 07-11-12.
65. ^ Foley, John. "Private Clouds Take Shape" (http://www.informationweek.com/news/services/business/showArticle.jhtml?articleID=209904474). InformationWeek. Retrieved 2010-08-22.
66. ^ Haff, Gordon (2009-01-27). "Just don't call them private clouds" (http://news.cnet.com/8301-13556_3-10150841-61.html). CNET News. Retrieved 2010-08-22.
67. ^ "There's No Such Thing As A Private Cloud" (http://www.informationweek.com/cloud-computing/theres-no-such-thing-as-a-private-cloud/229207922). InformationWeek. 2010-06-30. Retrieved 2010-08-22.
68. ^ Metzler, Jim; Taylor, Steve. (2010-08-23) "Cloud computing: Reality vs. fiction," Network World. [1] (http://www.networkworld.com/newsletters/frame/2010/082310wan1.html)
69. ^ Rouse, Margaret. "Definition: Cloudbursting," May 2011. SearchCloudComputing.com. [2] (http://searchcloudcomputing.techtarget.com/definition/cloud-bursting)
70. ^ Vizard, Michael. "How Cloudbursting 'Rightsizes' the Data Center", (2012-06-21). Slashdot. [3] (http://slashdot.org/topic/datacenter/how-cloudbursting-rightsizes-the-data-center/)
71. ^ Stevens, Alan (June 29, 2011). "When hybrid clouds are a mixed blessing" (http://www.theregister.co.uk/2011/06/29/hybrid_cloud/). *The Register* (http://www.theregister.co.uk). Retrieved March 28, 2012.
72. ^ "Building GrepTheWeb in the Cloud, Part 1: Cloud Architectures" (http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1632&categoryID=100). Developer.amazonwebservices.com. Retrieved 2010-08-22.
73. ^ Bernstein, David; Ludvigson, Erik; Sankar, Krishna; Diamond, Steve; Morrow, Monique (2009-05-24),

"Blueprint for the Intercloud - Protocols and Formats for Cloud Computing Interoperability" (http://www2.computer.org/portal/web/csdl/doi/10.1109/ICIW.2009.55). *Blueprint for the Intercloud – Protocols and Formats for Cloud Computing Interoperability*. IEEE Computer Society. pp. 328–336. doi:10.1109/ICIW.2009.55 (http://dx.doi.org/10.1109%2FICIW.2009.55). ISBN 978-1-4244-3851-8.
74. ^ "Kevin Kelly: A Cloudbook for the Cloud" (http://www.kk.org/thetechnium/archives/2007/11/a_cloudbook_for.php). Kk.org. Retrieved 2010-08-22.
75. ^ "Intercloud is a global cloud of clouds" (http://samj.net/2009/06/intercloud-is-global-cloud-of-clouds.html). Samj.net. 2009-06-22. Retrieved 2010-08-22.
76. ^ "Vint Cerf: Despite Its Age, The Internet is Still Filled with Problems" (http://www.readwriteweb.com/archives/vint_cerf_despite_its_age_the.php?mtcCampaign=2765). Readwriteweb.com. Retrieved 2010-08-22.
77. ^ "SP360: Service Provider: From India to Intercloud" (http://blogs.cisco.com/sp/comments/from_india_to_intercloud/). Blogs.cisco.com. Retrieved 2010-08-22.
78. ^ Canada (2007-11-29). "Head iaaan the clouds? Welcome to the future" (http://www.theglobeandmail.com/servlet/story/LAC.20071129.TWLINKS29/TPStory/Business). Toronto: Theglobeandmail.com. Retrieved 2010-08-22.
79. ^ Bobby Johnston. Cloud computing is a trap, warns GNU founder Richard Stallman. The Guardian, 29 September 2008.
80. ^ http://www.morganstanley.com/views/perspectives/cloud_computing.pdf
81. ^ *Challenges & Opportunities for IT partners when transforming or creating a business in the Cloud*. compuBase consulting. 2012. p. 77.
82. ^ Cloud Computing Grows Up: Benefits Exceed Expectations According to Report. Press Release, May 21, 2013. [4] (http://online.wsj.com/article/PR-CO-20130521-906501.html?mod=googlenews_wsj)
83. ^ Cauley, Leslie (2006-05-11). "NSA has massive database of Americans' phone calls" (http://www.usatoday.com/news/washington/2006-05-10-nsa_x.htm). USATODAY.com. Retrieved 2010-08-22.
84. ^ "NSA taps in to user data of Facebook, Google and others, secret files reveal" (http://www.guardian.co.uk/world/2013/jun/06/us-tech-giants-nsa-data). Guardian News and Media. 2013-06-07. Retrieved 2013-06-07.
85. ^ Winkler, Vic (2011). *Securing the Cloud: Cloud Computer Security Techniques and Tactics* (http://www.elsevier.com/wps/find/bookdescription.cws_home/723529/description). Waltham, Massachusetts: Elsevier. p. 60. ISBN 978-1-59749-592-9.
86. ^ "Feature Guide: Amazon EC2 Availability Zones" (http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1347&categoryID=112). Amazon Web Services. Retrieved 2010-08-22.
87. ^ "Cloud Computing Privacy Concerns on Our Doorstep" (http://cacm.acm.org/magazines/2011/1/103200-cloud-computing-privacy-concerns-on-our-doorstep/fulltext).
88. ^ "FISMA compliance for federal cloud computing on the horizon in 2010" (http://searchcompliance.techtarget.com/news/article/0,289142,sid195_gci1377298,00.html). SearchCompliance.com. Retrieved 2010-08-22.
89. ^ "Google Apps and Government" (http://googleenterprise.blogspot.com/2009/09/google-apps-and-government.html). Official Google Enterprise Blog. 2009-09-15. Retrieved 2010-08-22.
90. ^ "Cloud Hosting is Secure for Take-off: Mosso Enables The Spreadsheet Store, an Online Merchant, to become PCI Compliant" (http://www.rackspace.com/cloud/blog/2009/03/05/cloud-hosting-is-secure-for-take-off-mosso-enables-the-spreadsheet-store-an-online-merchant-to-become-pci-compliant/). Rackspace. 2009-03-14. Retrieved 2010-08-22.
91. ^ "Amazon gets SAS 70 Type II audit stamp, but analysts not satisfied" (http://searchcloudcomputing.techtarget.com/news/article/0,289142,sid201_gci1374629,00.html). SearchCloudComputing.com. 2009-11-17. Retrieved 2010-08-22.
92. ^ "Assessing Cloud Computing Agreements and Controls" (http://wistechnology.com/articles/6954/). WTN News. Retrieved 2010-08-22.
93. ^ "Cloud Certification From Compliance Mandate to Competitive Differentiator" (http://www.youtube.com/watch?v=wYiFdnZAlNQ). Cloudcor. Retrieved 2011-09-20.
94. ^ "How the New EU Rules on Data Export Affect Companies in and Outside the EU | Dr. Thomas Helbing – Kanzlei für Datenschutz-, Online- und IT-Recht" (http://www.thomashelbing.com/en/how-new-eu-rules-data

Kanzlei für Datenschutz-, Online- und IT-Recht" (http://www.thomashelbing.com/en/how-new-eu-rules-data-export-affect-companies-and-outside-eu). Dr. Thomas Helbing. Retrieved 2010-08-22.
95. ^ "FedRAMP" (http://www.gsa.gov/portal/category/102371). *U.S. General Services Administration*. 2012-06-13. Retrieved 2012-06-17.
96. ^ ***a b*** Chambers, Don (July 2010). "Windows Azure: Using Windows Azure's Service Bus to Solve Data Security Issues]" (http://rebustechnologies.com/wp-content/uploads/2011/12/WindowsAzure.pdf). *Rebus Technologies*. Retrieved 2012-12-14.
97. ^ Cohn, Cindy; Samuels, Julie (31 October 2012). "Megaupload and the Government's Attack on Cloud Computing]" (https://www.eff.org/deeplinks/2012/10/governments-attack-cloud-computing). *Electronic Frontier Foundation*. Retrieved 2012-12-14.
98. ^ Maltais, Michelle (26 April 2012). "Who owns your stuff in the cloud?" (http://articles.latimes.com/2012/apr/26/business/la-fi-tech-savvy-cloud-services-20120426). *Los Angeles Times*. Retrieved 2012-12-14.
99. ^ ***a b c*** McKendrick, Joe. (2011-11-20) "Cloud Computing's Vendor Lock-In Problem: Why the Industry is Taking a Step Backward," Forbes.com [5] (http://www.forbes.com/sites/joemckendrick/2011/11/20/cloud-computings-vendor-lock-in-problem-why-the-industry-is-taking-a-step-backwards/)
100. ^ Hinkle, Mark. (2010-6-9) "Three cloud lock-in considerations", Zenoss Blog [6] (http://community.zenoss.org/blogs/zenossblog/2010/06/09/three-cloud-lock-in-considerations)
101. ^ Staten, James (2012-07-23). "Gelsinger brings the 'H' word to VMware". ZDNet. [7] (http://www.zdnet.com/gelsinger-brings-the-h-word-to-vmware-7000001416/)
102. ^ Vada, Eirik T. (2012-06-11) "Creating Flexible Heterogeneous Cloud Environments", page 5, Network and System Administration, Oslo University College [8] (https://www.duo.uio.no/bitstream/handle/123456789/34153/thesis.pdf?sequence=1)
103. ^ Geada, Dave. (June 2, 2011) "The case for the heterogeneous cloud," Cloud Computing Journal [9] (http://cloudcomputing.sys-con.com/node/1841850)
104. ^ Burns, Paul (2012-01-02). "Cloud Computing in 2012: What's Already Happening". Neovise.[10] (http://www.neovise.com/cloud-computing-in-2012-what-is-already-happening)
105. ^ ***a b c d*** Livenson, Ilja. Laure, Erwin. (2011) "Towards transparent integration of heterogeneous cloud storage platforms", pages 27–34, KTH Royal Institute of Technology, Stockholm, Sweden. [11] (http://dl.acm.org/citation.cfm?id=1996020)
106. ^ ***a b*** Gannes, Liz. GigaOm, "Structure 2010: Intel vs. the Homogeneous Cloud," June 24, 2010. [12] (http://gigaom.com/2010/06/24/structure-2010-intel-vs-the-homogeneous-cloud/)
107. ^ Jon Brodkin (July 28, 2008). "Open source fuels growth of cloud computing, software-as-a-service" (http://www.networkworld.com/news/2008/072808-open-source-cloud-computing.html). Network World. Retrieved 2012-12-14.
108. ^ "VMware Launches Open Source PaaS Cloud Foundry" (http://www.cmswire.com/cms/information-management/vmware-launches-open-source-paas-cloud-foundry-010941.php). Simpler Media Group, Inc. 2011-04-21. Retrieved 2012-12-14.
109. ^ "AGPL: Open Source Licensing in a Networked Age" (http://redmonk.com/sogrady/2009/04/15/open-source-licensing-in-a-networked-age/). Redmonk.com. 2009-04-15. Retrieved 2010-08-22.
110. ^ GoGrid Moves API Specification to Creative Commons (http://www.gogrid.com/company/press-releases/gogrid-moves-api-specification-to-creativecommons.php)
111. ^ "Eucalyptus Completes Amazon Web Services Specs with Latest Release" (http://ostatic.com/blog/eucalyptus-completes-amazon-web-services-specs-with-latest-release). Ostatic.com. Retrieved 2010-08-22.
112. ^ "OpenStack Foundation launches" (http://www.infoworld.com/d/cloud-computing/openstack-foundation-launches-202694/). Infoworld.com. 2012-09-19. Retrieved 2012-17-11.
113. ^ "Did OpenStack Let VMware Into The Henhouse?" (http://www.informationweek.com/development/open-source/did-openstack-let-vmware-into-the-henhou/240009337). Informationweek.com. 2012-10-19. Retrieved 2012-17-11.
114. ^ ***a b*** Zissis, Dimitrios; Lekkas (2010). "Addressing cloud computing security issues" (http://www.sciencedirect.com/science/article/pii/S0167739X10002554). *Future Generation Computer Systems* **28** (3): 583. doi:10.1016/j.future.2010.12.006 (http://dx.doi.org/10.1016%2Fj.future.2010.12.006).
115. ^ Winkler, Vic (2011). *Securing the Cloud: Cloud Computer Security Techniques and Tactics*

(http://www.elsevier.com/wps/find/bookdescription.cws_home/723529/description#description). Waltham, MA USA: Syngress. pp. 187, 189. ISBN 978-1-59749-592-9.
116. ^ "Are security issues delaying adoption of cloud computing?" (http://www.networkworld.com/news/2009/042709-burning-security-cloud-computing.html). Network World. Retrieved 2010-08-22.
117. ^ "Security of virtualization, cloud computing divides IT and security pros" (http://www.networkworld.com/news/2010/022210-virtualization-cloud-security-debate.html). Network World. 2010-02-22. Retrieved 2010-08-22.
118. ^ Armbrust, M; Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Zaharia, (2010). "A view of cloud computing". *Communication of the ACM* **53** (4): 50–58. doi:10.1145/1721654.1721672 (http://dx.doi.org/10.1145%2F1721654.1721672).
119. ^ Anthens, G (2010). "Security in the cloud". *Communications of the ACM* **53** (11): 16. doi:10.1145/1839676.1839683 (http://dx.doi.org/10.1145%2F1839676.1839683).
120. ^ James Urquhart (January 7, 2010). "Cloud computing's green paradox" (http://news.cnet.com/8301-19413_3-10428065-240.html). CNET News. Retrieved March 12, 2010. "... there is some significant evidence that the cloud is encouraging more compute consumption"
121. ^ Finland – First Choice for Siting Your Cloud Computing Data Center. (http://www.fincloud.freehostingcloud.com/). Retrieved 4 August 2010.
122. ^ Swiss Carbon-Neutral Servers Hit the Cloud. (http://www.greenbiz.com/news/2010/06/30/swiss-carbon-neutral-servers-hit-cloud). Retrieved 4 August 2010.
123. ^ Berl, Andreas, et al., Energy-Efficient Cloud Computing (http://comjnl.oxfordjournals.org/content/53/7/1045.short), The Computer Journal, 2010.
124. ^ Farrahi Moghaddam, Fereydoun, et al., Low Carbon Virtual Private Clouds (http://ieeexplore.ieee.org/search/srchabstract.jsp?tp=&arnumber=6008718), IEEE Cloud 2011.
125. ^ Alpeyev, Pavel (2011-05-14). "Amazon.com Server Said to Have Been Used in Sony Attack" (http://www.bloomberg.com/news/2011-05-13/sony-network-said-to-have-been-invaded-by-hackers-using-amazon-com-server.html). Bloomberg. Retrieved 2011-08-20.
126. ^ Goodin, Dan (2011-05-14). "PlayStation Network hack launched from Amazon EC2" (http://www.theregister.co.uk/2011/05/14/playstation_network_attack_from_amazon/). The Register. Retrieved 2012-05-18.
127. ^ Hsu, Wen-Hsi L., "Conceptual Framework of Cloud Computing Governance Model - An Education Perspective", IEEE Technology and Engineering Education (ITEE), Vol 7, No 2 (2012) [13] (http://www.ewh.ieee.org/soc/e/sac/itee/index.php/meem/article/view/240)
128. ^ Stackpole, Beth, "Governance Meets Cloud: Top Misconceptions", InformationWeek, 7 May 2012 [14] (http://www.informationweek.com/cloud-computing/infrastructure/governance-meets-cloud-top-misconception/232901483)
129. ^ Joha, A and M. Janssen (2012) "Transformation to Cloud Services Sourcing: Required IT Governance Capabilities", ICST Transactions on e-Business 12(7-9) [15] (http://eudl.eu/pdf/10.4108/eb.2012.07-09.e4)
130. ^ ***a b*** Gardner, Jake (2013-03-28). "Beware: 7 Sins of Cloud Computing" (http://www.wired.com/insights/2013/01/beware-7-sins-of-cloud-computing). Wired.com. Retrieved 2013-06-20.
131. ^ S. Stonham and S. Nahalkova (2012) "What is the Cloud and how can it help my business?" [16] (http://www.ovasto.com/2013/01/what-is-the-cloud-how-can-the-cloud-help-my-business/)
132. ^ S. Stonham and S. Nahalkova (2012), Whitepaper "Tomorrow Belongs to the Agile (PDF)" [17] (http://www.ovasto.com/full-service-business-marketing-consultancy/strategic-agility-and-the-cloud/)
133. ^ George Kousiouris, Tommaso Cucinotta, Theodora Varvarigou, "The Effects of Scheduling, Workload Type and Consolidation Scenarios on Virtual Machine Performance and their Prediction through Optimized Artificial Neural Networks"[18] (http://users.ntua.gr/gkousiou/publications/JSS_Kousiouris.pdf) , The Journal of Systems and Software (2011),Volume 84, Issue 8, August 2011, pp. 1270-1291, Elsevier, doi:10.1016/j.jss.2011.04.013.
134. ^ "Cloud Net Directory. Retrieved 2010-03-01" (http://www.cloudbook.net/directories/research-clouds). Cloudbook.net. Retrieved 2010-08-22.
135. ^ "– National Science Foundation (NSF) News – National Science Foundation Awards Millions to Fourteen Universities for Cloud Computing Research – US National Science Foun" (http://www.nsf.gov/news/news_summ.jsp?cntn_id=114686). Nsf.gov. Retrieved 2011-08-20.

(http://www.nist.gov/news/news_sumfm.jsp?cntnt_id=114686). Nist.gov. Retrieved 2011-08-20.
136. ^ Rich Miller (2008-05-02). "IBM, Google Team on an Enterprise Cloud" (http://www.datacenterknowledge.com/archives/2008/05/02/ibm-google-team-on-an-enterprise-cloud/). DataCenterKnowledge.com. Retrieved 2010-08-22.
137. ^ "StACC - Collaborative Research in Cloud Computing" (http://www.cs.st-andrews.ac.uk/stacc). University of St Andrews department of Computer Science. Retrieved 2012-06-17.
138. ^ "Trustworthy Clouds: Privacy and Resilience for Internet-scale Critical Infrastructure" (http://www.tclouds-project.eu). Retrieved 2012-06-17.
139. ^ *a b* Ko, Ryan K. L.; Jagadpramana, Peter; Lee, Bu Sung (2011). "Flogger: A File-centric Logger for Monitoring File Access and Transfers within Cloud Computing Environments" (http://www.hpl.hp.com/techreports/2011/HPL-2011-119.pdf). *Proceedings of the 10th IEEE International Conference on Trust, Security and Privacy of Computing and Communications (TrustCom-11)*: 765. doi:10.1109/TrustCom.2011.100 (http://dx.doi.org/10.1109%2FTrustCom.2011.100). ISBN 978-1-4577-2135-9.
140. ^ Ko, Ryan K. L.; Jagadpramana, Peter; Mowbray, Miranda; Pearson, Siani; Kirchberg, Markus; Liang, Qianhui; Lee, Bu Sung (2011). "TrustCloud: A Framework for Accountability and Trust in Cloud Computing" (http://www.hpl.hp.com/techreports/2011/HPL-2011-38.pdf). *Proceedings of the 2nd IEEE Cloud Forum for Practitioners (IEEE ICFP 2011), Washington DC, USA, July 7–8, 2011*.
141. ^ Ko, Ryan K. L. Ko; Kirchberg, Markus; Lee, Bu Sung (2011). "From System-Centric Logging to Data-Centric Logging - Accountability, Trust and Security in Cloud Computing" (http://www.hpl.hp.com/people/ryan_ko/RKo-DSR2011-Data_Centric_Logging.pdf). *Proceedings of the 1st Defence, Science and Research Conference 2011 - Symposium on Cyber Terrorism, IEEE Computer Society, 3–4 August 2011, Singapore*.
142. ^ "UTM/UPES-IBM India Collaboration" (http://www.youtube.com/watch?v=bDVKQKBN8XY). 2011.
143. ^ "Publication Download" (http://www.tiaonline.org/market_intelligence/publication_download.cfm?file=TIA_Cloud_Computing_White_Paper). Tiaonline.org. Retrieved 2011-12-02.
144. ^ "Testbeds for cloud experimentation and testing" (http://www.bonfire-project.eu). Retrieved 2013-04-09.

External links

- The NIST Definition of Cloud Computing (http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf). Peter Mell and Timothy Grance, NIST Special Publication 800-145 (September 2011). National Institute of Standards and Technology, U.S. Department of Commerce.
- Guidelines on Security and Privacy in Public Cloud Computing (http://nvlpubs.nist.gov/nistpubs/sp/2011/sp800-144.pdf). Wayne Jansen and Timothy Grance, NIST Special Publication 800-144 (December 2011). National Institute of Standards and Technology, U.S. Department of Commerce.
- Cloud Computing - Benefits, risks and recommendation for information security (http://www.enisa.europa.eu/activities/risk-management/files/deliverables/cloud-computing-risk-assessment). Daniele Catteddu and Giles Hobben, European Network and Information Security Agency 2009.
- Fighting cyber crime and protecting privacy in the cloud. European Parliament - Directorate-General for Internal Policies. 2012 (http://www.europarl.europa.eu/committees/en/studiesdownload.html?languageDocument=EN&file=79050)
- Cloud Computing: What are the Security Implications?: Hearing before the Subcommittee on Cybersecurity, Infrastructure Protection, and Security Technologies of the Committee on Homeland Security, House of Representatives, One Hundred Twelfth Congress, First Session, October 6, 2011 (http://purl.fdlp.gov/GPO/gpo32975)
- Cloud Computing represents both a significant opportunity and a potential challenge (http://ptop.co.uk/solutions-and-technology/cloud-computing/)
- Cloud and Datacenter Solution Hub (http://technet.microsoft.com/en-us/cloud/private-cloud) on Microsoft

TechNet

The Secret to Creating Value from Web Services Today: Start Simply

By John Hagel, John Seely Brown, and Dennis Layton-Rodin

As enthusiasm around Web services gains momentum, many CIO's are reacting with skepticism and concern. Whether it is lack of confidence in the ability of partners to deliver mission critical services, concern about allowing entities outside the enterprise to perforate the firewall, or reluctance to unleash entrepreneurial initiatives within the IT department, CIO's see Web services as a potential minefield.

However, Web services offer real business benefits today so CIO's should resist the urge to build roadblocks. Instead, CIO's need to find ways to gain experience and reap some of the benefits of this new technology without exposing the enterprise to undue risk. The most promising path in Web services is "Keep It Simple, Keep it Incremental, and Learn, Learn, Learn". By simple, we mean working with simply structured data, simple protocols and simple business processes. By incremental, we mean proceeding in small steps so you (and your partners) can learn what works best and build confidence in the technology.

Keep It Simple

Simple data: By focusing on specific event based information that directly impacts the actions of the enterprise, CIO's can help to simplify their company's approach to data sharing. Dell Computer illustrates this opportunity. Dell is seeking to lower the inventory in their supply chain through more focused information sharing with their partners.

Dell seeks to fulfill orders within 5 days of receipt of order, but it takes their suppliers up to 45 days to fulfill orders. For this reason, Dell used to carry to inventories of up to 30 hours in their factories. Since their factories and their partners all operated on disparate systems, information sharing was very labor intensive. Rather trying to achieve transparency across their supply chain partners by imposing a common technology platform (like Cisco is attempting), they sought much more limited visibility through the sharing of simple event acknowledgements (e.g., product shipped on time).

By using Web services to automatically process simple acknowledgements of met commitments across many disparate systems, Dell can focus its staff on exception handling to work around disruptions in the supply chain before they become problems. Using Web services, Dell's supply chain partners are able to avoid expensive new IT investments. Dell has been able to reduce its raw materials inventory to just 3- 5 hours and it is now working on a second stage designed to reduce inventory among its supply chain partners as well.

Simple protocols: The key challenge in integrating across disparate systems is getting divergent applications -- let alone divergent frameworks -- to talk with each other. An alternative approach is to use very simple protocols (e.g., SOAP, FTP) to just move the data from its source to its target. Once the data is at its target, simple scripts can be crafted to insert the data into the applications for use. By focusing on transferring specific data as opposed to invoking applications, it is possible to bridge different frameworks and allow incompatible environments to work together.

Simple business processes: Dell succeeded because it worked hard to reduce its process to the lowest common denominator. They realized that "supply chain orchestration" is a complex process, but much of that complex process can be reduced to a series of very simple, almost binary, communications. At the end of the day, partners need to share the information a process conveys, not the process itself.

Keep It Incremental

CIO's can manage risk by staging the implementation of Web services in a series of small, low risk steps. This staging can occur in multiple dimensions.

Business partners: Start with a limited number of well-established business partners where you have already built strong trust-based relationships and a deep understanding of each other's business. In Dell's case, they started with less than a dozen vendor-managed hubs – specialized third party logistics providers that coordinated shipments from hundreds of suppliers. As Dell gained experience with its approach to event management using Web services, it broadened its focus to include its suppliers, starting with primary suppliers where Dell had established strong relationships over many years. As a company gains more experience with Web services, it can expand the number and diversity of business partners involved, even bringing in new business partners without any pre-existing relationships.

Level of specification: In business relationships involving high value processes, companies have traditionally sought to manage risk by negotiating detailed contracts that specify in great detail the activities to be performed by each partner. While such high levels of specification help to reduce some forms of risk, they actually introduce other risks. In particular, they can reduce flexibility to adapt to unforeseen changes in market conditions.

Web services technology creates the potential for more flexibility in collaborative business processes. To exploit this potential, companies need to shift to from approaches involving high specification of activities to lower specification approaches relying on other, more flexible, ways (such as incentives and selective information visibility) to achieve desired business results.

Once again, the answer is not to move overnight from high specification approaches to low specification approaches. Instead, stage the transition by selectively reducing the

range of specification and implementing alternative approaches to coordination. By beginning with trusted business partners, a company can gain experience with these approaches before applying them to newer business partners.

Amount of value: CIO's can also manage risk by initially focusing on individual business activities that are relatively low value and expanding over time to include much higher value activities. For example, many financial services firms are starting to use Web services to distribute content like investment analyst reports to their clients. The business risk in such applications is relatively low, but it can help to increase client satisfaction through more convenient access to information. As these firms gain more experience in the application of Web services technology, they can begin to expand the range of applications to higher value (and higher risk) activities like dissemination of investment portfolio data and processing of client transactions.

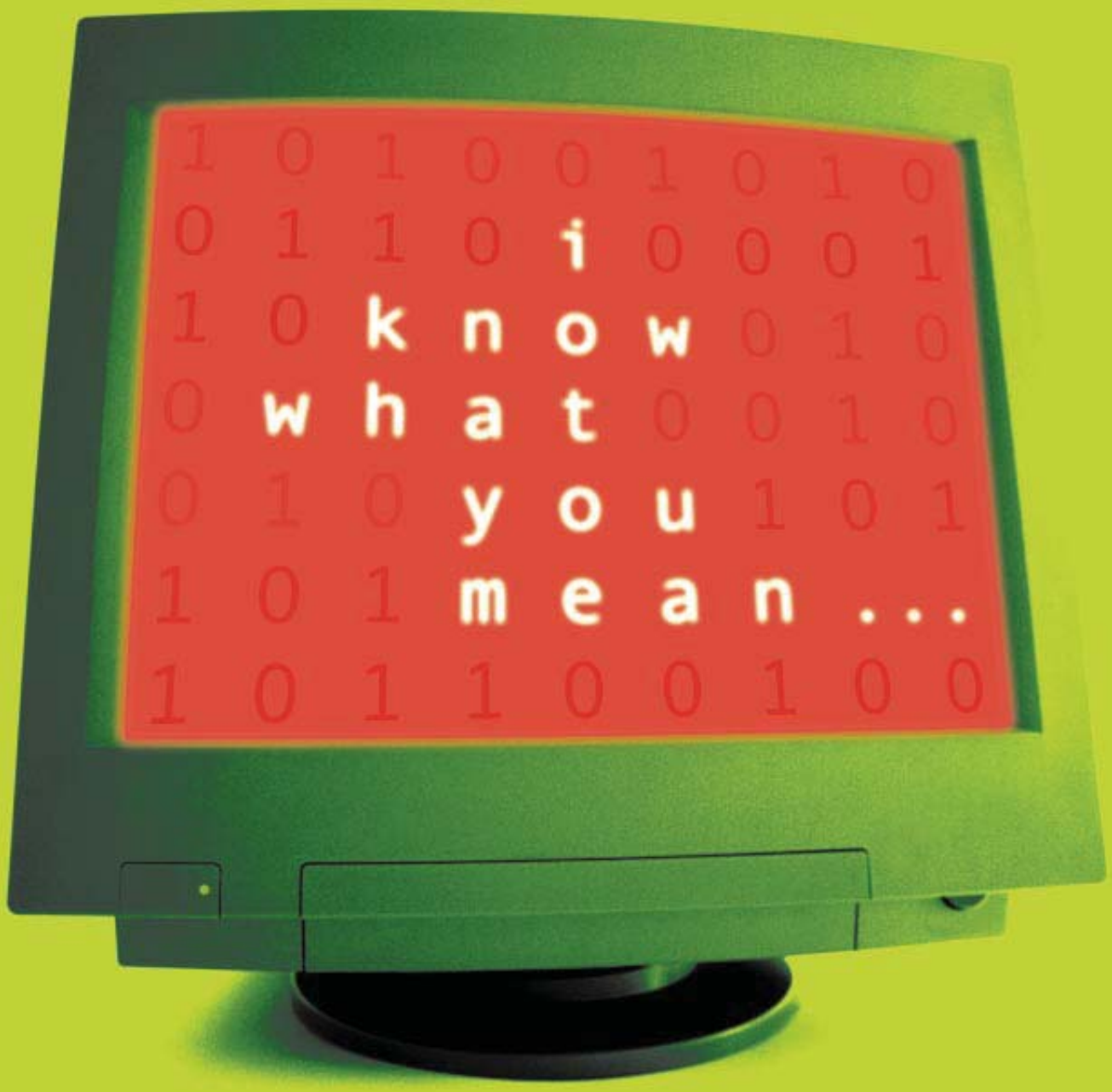
Learn, Learn, Learn

Ultimately, the key to reducing risk and increasing business value is to learn from early experience and deepen skills in the application and operation of Web services technology. As skills improve, the technology can be applied in a broader range of business environments while still controlling risk.

Learning does not happen automatically. CIO's need to be thoughtful about what can be learned from early applications of Web services technology and design appropriate information feedback loops to support the learning process. The good news is that, by automating connections across applications, Web services technology can generate very helpful data about the performance of the technology. CIO's need to capture this data and transform it into useful information to support learning. By doing this, CIO's can significantly accelerate the incremental approach outlined above and more rapidly address opportunities to generate significant value without exposing the enterprise to undue risk. Start simply, but start – the hardest way to learn is to not do anything.

John Hagel III is an independent management consultant who work focuses on the intersection of business strategy and technology. His most recent book, Out of the Box: Strategies for Achieving Profits Today and Growth Tomorrow, was published by Harvard Business School Press in October 2002. He can be reached through his website www.johnhagel.com or by e-mail at john@johnhagel.com

John Seely Brown was the director of Xerox PARC until 2000. He continues his personal research into digital culture, learning and Web services. His most recent book (co-authored with Paul Duguid) is The Social Life of Information. He can be reached by e-mail at jsb@parc.com.



THE SEMANTIC WEB

A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

by
TIM BERNERS-LEE,
JAMES HENDLER and
ORA LASSILA

PHOTOILLUSTRATIONS BY MIGUEL SALMERON

*The entertainment system was belting out the Beatles' "We Can Work It Out" when the phone rang. When Pete answered, his phone turned the sound down by sending a message to all the other **local** devices that had a **volume control**. His sister, Lucy, was on the line from the doctor's office: "Mom needs to see a specialist and then has to have*

a series of physical therapy sessions. Biweekly or something. I'm going to have my agent set up the appointments." Pete immediately agreed to share the chauffeuring.

At the doctor's office, Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's ***prescribed treatment*** from the doctor's agent, looked up several lists of ***providers,*** and checked for the ones ***in-plan*** for Mom's insurance within a ***20-mile radius*** of her ***home*** and with a ***rating*** of ***excellent*** or ***very good*** on trusted rating services. It then began trying to find a match between available ***appointment times*** (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules. (The emphasized keywords indicate terms whose semantics, or meaning, were defined for the agent through the Semantic Web.)

In a few minutes the agent presented them with a plan. Pete didn't like it—University Hospital was all the way across town from Mom's place, and he'd be driving back in the middle of rush hour. He set his own agent to redo the search with stricter preferences about ***location*** and ***time***. Lucy's agent, having ***complete trust*** in Pete's agent in the context of the present task, automatically assisted by supplying access certificates and shortcuts to the data it had already sorted through.

Almost instantly the new plan was presented: a much closer clinic and earlier times—but there were two warning notes. First, Pete would have to reschedule a couple of his ***less important*** appointments. He checked what they were—not a problem. The other was something about the insurance company's list failing to include this provider under ***physical therapists:*** "Service type and insurance plan status securely verified by other means," the agent reassured him. "(Details?)"

Lucy registered her assent at about the same moment Pete was muttering, "Spare me the details," and it was all set. (Of course, Pete couldn't resist the details and later that night had his agent explain how it had found that provider even though it wasn't on the proper list.)

Expressing Meaning

PETE AND LUCY could use their agents to carry out all these tasks thanks not to the World Wide Web of today but rather the Semantic Web that it will evolve into tomorrow. Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully. Computers can adeptly parse Web pages for layout and routine processing—here a header, there a link to another page—but in general, computers have no reliable way to process the semantics: this is the home page of the Hartman and Strauss Physio Clinic, this link goes to Dr. Hartman's curriculum vitae.

The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. Such an agent coming to the clinic's Web page will know not just that the page has keywords such as "treatment, medicine, physical, therapy"

Overview / *Semantic Web*

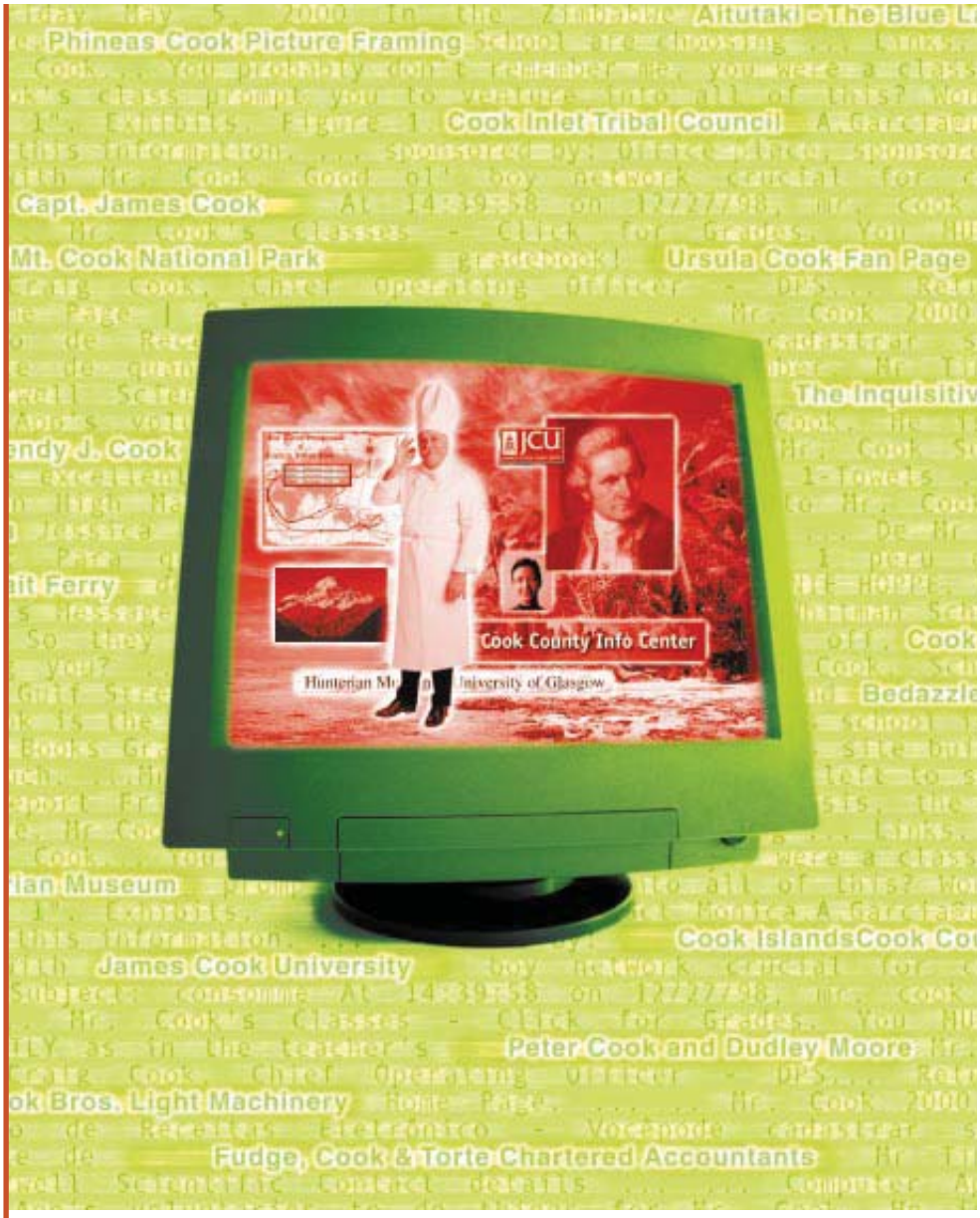
- To date, the World Wide Web has developed most rapidly as a medium of documents for people rather than of information that can be manipulated automatically. By augmenting Web pages with data targeted at computers and by adding documents solely for computers, we will transform the Web into the Semantic Web.
- Computers will find the meaning of semantic data by following hyperlinks to definitions of key terms and rules for reasoning about them logically. The resulting infrastructure will spur the development of automated Web services such as highly functional agents.
- Ordinary users will compose Semantic Web pages and add new definitions and rules using off-the-shelf software that will assist with semantic markup.

(as might be encoded today) but also that Dr. Hartman ***works*** at this ***clinic*** on ***Mondays, Wednesdays*** and ***Fridays*** and that the script takes a ***date range*** in ***yyyy-mm-dd format*** and returns ***appointment times***. And it will "know" all this without needing artificial intelligence on the scale of *2001*'s Hal or *Star Wars*'s C-3PO. Instead these semantics were encoded into the Web page when the clinic's office manager (who never took Comp Sci 101) massaged it into shape using off-the-shelf software for writing Semantic Web pages along with resources listed on the Physical Therapy Association's site.

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and "understand" the data that they merely display at present.

The essential property of the World Wide Web is its universality. The power of a hypertext link is that "anything can link to anything." Web technology, therefore, must not discriminate between the scribbled draft and the polished performance, between commercial and academic information, or among cultures, languages, media and so on. Information varies along many axes. One of these is the difference between information produced primarily for human consumption and that produced mainly for machines. At one end of the scale we have everything from the five-second TV commercial to poetry. At the other end we have databases, programs and sensor output. To date, the Web has developed most rapidly as a medium of documents for people rather than for data and information that can be processed automatically. The Semantic Web aims to make up for this.

Like the Internet, the Semantic Web will be as decentralized as possible. Such Web-like systems generate a lot of excitement at every level, from major corporation to individual user, and provide benefits that are hard or impossible to predict in advance. Decentralization requires compromises: the Web had to throw away the ideal of total consistency of all of its interconnections, ushering in the infamous message "Error 404: Not Found" but allowing unchecked exponential growth.



WEB SEARCHES TODAY typically turn up innumerable completely irrelevant "hits," requiring much manual filtering by the user. If you search using the keyword "cook," for example, the computer has no way of knowing whether you are looking for a chef, information about how to cook something, or simply a place, person, business or some other entity with "cook" in its name. The problem is that the word "cook" has no meaning, or semantic content, to the computer.

Knowledge Representation

FOR THE SEMANTIC WEB to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. Artificial-intelligence researchers have studied such systems since long before the Web was developed. Knowledge representation, as this technology is often called, is currently in a state comparable to that of hypertext before the advent of the Web: it is clearly a good idea, and some very nice demonstrations exist, but it has not yet changed the world. It contains the seeds of important applications, but to realize its full potential it must be linked into a single global system.

Traditional knowledge-representation systems typically have been centralized, requiring everyone to share exactly the same definition of common concepts

Glossary

HTML: Hypertext Markup Language. The language used to encode formatting, links and other features on Web pages. Uses standardized "tags" such as <H1> and <BODY> whose meaning and interpretation is set universally by the World Wide Web Consortium.

XML: eXtensible Markup Language. A markup language like HTML that lets individuals define and use their own tags. XML has no built-in mechanism to convey the meaning of the user's new tags to other users.

RESOURCE: Web jargon for any entity. Includes Web pages, parts of a Web page, devices, people and more.

URL: Uniform Resource Locator. The familiar codes (such as http://www.sciam.com/index.html) that are used in hyperlinks.

URI: Universal Resource Identifier. URLs are the most familiar type of URI. A URI defines or specifies an entity, not necessarily by naming its location on the Web.

RDF: Resource Description Framework. A scheme for defining information on the Web. RDF provides the technology for expressing the meaning of terms and concepts in a form that computers can readily process. RDF can use XML for its syntax and URIs to specify entities, concepts, properties and relations.

ONTOLOGIES: Collections of statements written in a language such as RDF that define the relations between concepts and specify logical rules for reasoning about them. Computers will "understand" the meaning of semantic data on a Web page by following links to specified ontologies.

AGENT: A piece of software that runs without direct human control or constant supervision to accomplish goals provided by a user. Agents typically collect, filter and process information found on the Web, sometimes with the help of other agents.

SERVICE DISCOVERY: The process of locating an agent or automated Web-based service that will perform a required function. Semantics will enable agents to describe to one another precisely what function they carry out and what input data are needed.

such as "parent" or "vehicle." But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable.

Moreover, these systems usually carefully limit the questions that can be asked so that the computer can answer reliably—or answer at all. The problem is reminiscent of Gödel's theorem from mathematics: any system that is complex enough to be useful also encompasses unanswerable questions, much like sophisticated versions of the basic paradox "This sentence is false." To avoid such problems, traditional knowledge-representation systems generally each had their own narrow and idiosyncratic set of rules for making inferences about their data. For example, a genealogy system, acting on a database of family trees, might include the rule "a wife of an uncle is an aunt." Even if the data could be transferred from one system to another, the rules, existing in a completely different form, usually could not.

Semantic Web researchers, in contrast, accept that paradoxes and unanswerable questions are a price that must be paid to achieve versatility. We make the language for the rules as expressive as needed to allow the Web to reason as widely as desired. This philosophy is similar to that of the conventional Web: early in the Web's development, detractors pointed out that it could never be a well-organized library; without a central database and tree structure, one would never be sure of finding everything. They were right. But the expressive power of the system made vast amounts of information available, and search engines (which would have seemed quite impractical a decade ago) now produce remarkably complete indices of a lot of the material out there.

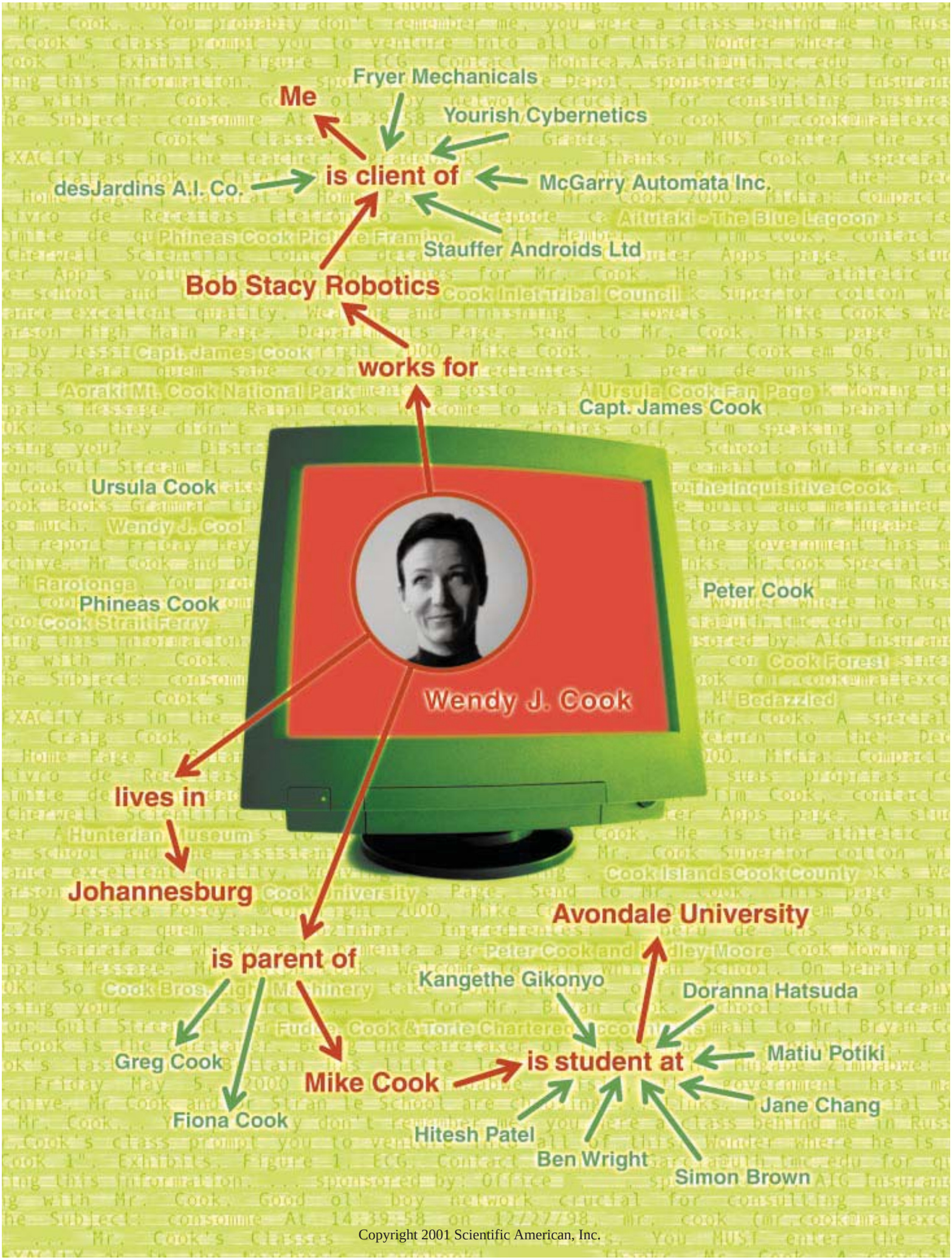
The challenge of the Semantic Web, therefore, is to provide a language that expresses both data and rules for reasoning about the data and that allows rules from any existing knowledge-representation system to be exported onto the Web.

Adding logic to the Web—the means to use rules to make inferences, choose courses of action and answer questions—is the task before the Semantic Web community at the moment. A mixture of mathematical and engineering decisions complicate this task. The logic must be powerful enough to describe complex properties of objects but not so powerful that agents can be tricked by being asked to consider a paradox. Fortunately, a large majority of the information we want to express is along the lines of "a hex-head bolt is a type of machine bolt," which is readily written in existing languages with a little extra vocabulary.

Two important technologies for developing the Semantic Web are already in place: eXtensible Markup Language (XML) and the Resource Description Framework (RDF). XML lets everyone create their own tags—hidden labels such as <zip code> or <alma mater> that annotate Web pages or sections of text on a page. Scripts, or programs, can make use of these tags in sophisticated ways, but the script writer has to know what the page writer uses each tag for. In short, XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean [see "XML and the Second-Generation Web," by Jon Bosak and Tim Bray; Scientific American, May 1999].

Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence. These triples

ELABORATE, PRECISE AUTOMATED SEARCHES will be possible when semantics are widespread on the Web. Here a search program correctly locates a person based on an assortment of partially remembered knowledge: her last name is "Cook," she works for a company on your client list, and she has a son attending your alma mater, Avondale University. The correct combination of that information does not reside on a single Web page, but semantics make it easier for a program to discern the elements on various pages, understand relations such as "Mike Cook is a child of Wendy Cook" and piece them together reliably. More generally, semantics will enable complicated processes and transactions to be carried out automatically.



can be written using XML tags. In RDF, a document makes assertions that particular things (people, Web pages or whatever) have properties (such as "is a sister of," "is the author of") with certain values (another person, another Web page). This structure turns out to be a natural way to describe the vast majority of the data processed by machines. Subject and object are each identified by a Universal Resource Identifier (URI), just as used in a link on a Web page. (URLs, Uniform Resource Locators, are the most common type of URI.) The verbs are also identified by URIs, which enables anyone to define a new concept, a new verb, just by defining a URI for it somewhere on the Web.

Human language thrives when using the same term to mean somewhat different things, but automation does not. Imagine that I hire a clown messenger service to deliver balloons to my customers on their birthdays. Unfortunately, the service transfers the addresses from my database to its database, not knowing that the "addresses" in mine are where bills are sent and that many of them are post office boxes. My hired clowns end up entertaining a number of postal workers—not necessarily a bad thing but certainly not the intended effect. Using a different URI for each specific concept solves that problem. An address that is a mailing address can be distinguished from one that is a street address, and both can be distinguished from an address that is a speech.

The triples of RDF form webs of information about related things. Because RDF uses URIs to encode this information in a document, the URIs ensure that concepts are not just words in a document but are tied to a unique definition that everyone can find on the Web. For example, imagine that we have access to a variety of databases with information about people, including their addresses. If we want to find people living in a specific zip code, we need to know which fields in each database represent names and which represent zip codes. RDF can specify that "(field 5 in database A) (is a field of type) (zip code)," using URIs rather than phrases for each term.

Ontologies

OF COURSE, THIS IS NOT the end of the story, because two databases may use different identifiers for what is in fact the same concept, such as ***zip code***. A program that wants to compare or combine information across the two databases has to know that these two terms are being used to mean the same thing. Ideally, the program must have a way to discover such common meanings for whatever databases it encounters.

A solution to this problem is provided by the third basic component of the Semantic Web, collections of information called ontologies. In philosophy, an ontology is a theory about the nature of existence, of what types of things exist; ontology as a discipline studies such theories. Artificial-intelligence and Web researchers have co-opted the term for their own jargon, and for them an ontology is a document or file that formally defines the relations among terms. The most typical kind of ontology for the Web has a taxonomy and a set of inference rules.

The taxonomy defines classes of objects and relations among them. For example, an ***address*** may be defined as a type of ***location,*** and ***city codes*** may be defined to apply only to ***locations,*** and so on. Classes, subclasses and relations among entities are a very powerful tool for Web use. We can express a large number of relations among entities by assigning properties to classes and allowing subclasses to inherit such properties. If ***city codes*** must be of type ***city*** and cities generally have Web sites, we can discuss the Web site associated with a ***city code*** even if no database links a city code directly to a Web site.

Inference rules in ontologies supply further power. An ontology may express the rule "If a city code is associated with a state code, and an address uses that city code, then that address has the associated state code." A program could then readily deduce, for instance, that a Cornell University address, being in Ithaca, must be in New York State, which is in the U.S., and therefore should be formatted to U.S. standards. The computer doesn't truly "understand" any of this information, but it can now manipulate the terms much more effectively in ways that are useful and meaningful to the human user.

With ontology pages on the Web, solutions to terminology (and other) problems begin to emerge. The meaning of terms or XML codes used on a Web page can be defined by pointers from the page to an ontology. Of course, the same problems as before now arise if I point to an

The Semantic Web will enable machines to COMPREHEND semantic documents and data, not human speech and writings.

THE AUTHORS

TIM BERNERS-LEE, JAMES HENDLER and *ORA LASSILA* are individually and collectively obsessed with the potential of Semantic Web technology. Berners-Lee is director of the World Wide Web Consortium (W3C) and a researcher at the Laboratory for Computer Science at the Massachusetts Institute of Technology. When he invented the Web in 1989, he intended it to carry more semantics than became common practice. Hendler is professor of computer science at the University of Maryland at College Park, where he has been doing research on knowledge representation in a Web context for a number of years. He and his graduate research group developed SHOE, the first Web-based knowledge representation language to demonstrate many of the agent capabilities described in this article. Hendler is also responsible for agent-based computing research at the Defense Advanced Research Projects Agency (DARPA) in Arlington, Va. Lassila is a research fellow at the Nokia Research Center in Boston, chief scientist of Nokia Venture Partners and a member of the W3C Advisory Board. Frustrated with the difficulty of building agents and automating tasks on the Web, he co-authored W3C's RDF specification, which serves as the foundation for many current Semantic Web efforts.

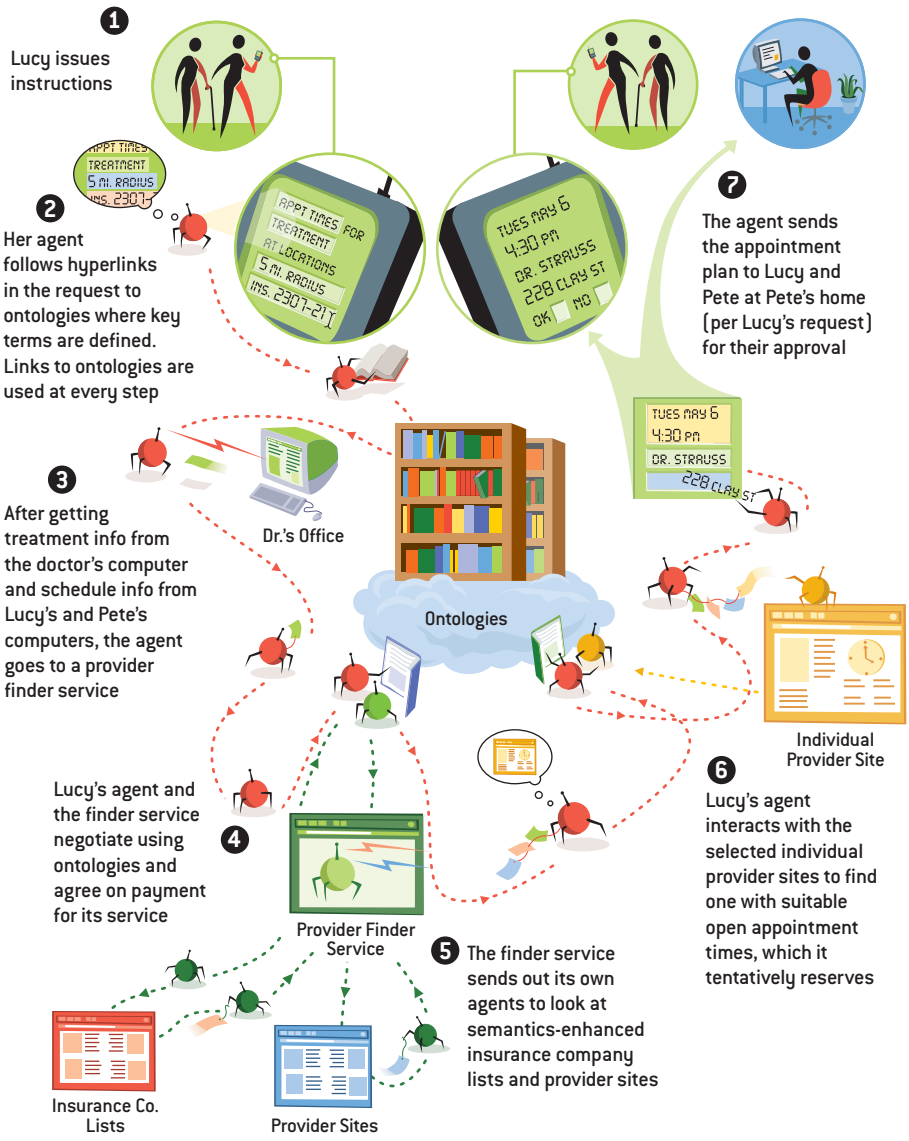
ontology that defines ***addresses*** as containing a ***zip code*** and you point to one that uses ***postal code***. This kind of confusion can be resolved if ontologies (or other Web services) provide equivalence relations: one or both of our ontologies may contain the information that my ***zip code*** is equivalent to your ***postal code***.

Our scheme for sending in the clowns to entertain my customers is partially solved when the two databases point to different definitions of ***address***. The program, using distinct URIs for different concepts of ***address***, will not confuse them and in fact will need to discover that the concepts are related at all. The program could then use a service that takes a list of postal ***addresses*** (defined in the first ontology) and converts it into a list of physical ***addresses*** (the second ontology) by recognizing and removing post office boxes and other unsuitable addresses. The structure and semantics provided by ontologies make it easier for an entrepreneur to provide such a service and can make its use completely transparent.

Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches—the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords. More advanced applications will use ontologies to relate the information on a page to the associated knowledge structures and inference rules. An example of a page marked up for such use is online at http://www.cs.umd.edu/~hendler. If you send your Web browser to that page, you will see the normal Web page entitled "Dr. James A. Hendler." As a human, you can readily find the link to a short biographical note and read there that Hendler received his Ph.D. from Brown University. A computer program trying to find such information, however, would have to be very complex to guess that this information might be in a biography and to understand the English language used there.

For computers, the page is linked to an ontology page that defines information about computer science departments. For instance, professors work at universities and they generally have doctorates. Further markup on the page (not displayed by the typical Web browser) uses the ontology's concepts to specify that Hendler received his Ph.D. from the entity described at the URI http://www.brown.edu/—the Web page for Brown. Computers can also find that Hendler is a member of a particular research project, has a particular e-mail address, and so on. All that information is readily processed by a computer and could be used to answer queries (such as where Dr. Hendler received his degree) that currently would require a human to sift through the content of various pages turned up by a search engine.

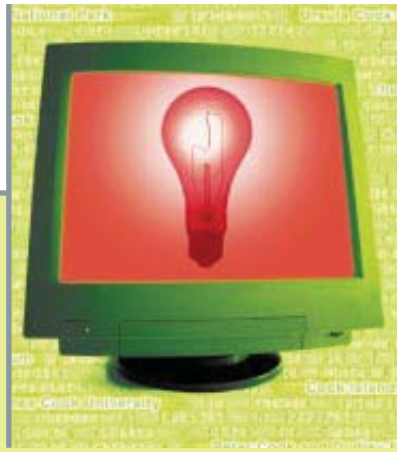
In addition, this markup makes it much easier to develop programs that can tackle complicated questions whose answers do not reside on a single Web page. Suppose you wish to find the Ms. Cook you met at a trade conference last year. You don't remember her first name, but you remember that she worked for one of your clients and that her son was a student at your alma mater. An intelligent search program can sift through all the pages of people whose name is



SOFTWARE AGENTS will be greatly facilitated by semantic content on the Web. In the depicted scenario, Lucy's agent tracks down a physical therapy clinic for her mother that meets a combination of criteria and has open appointment times that mesh with her and her brother Pete's schedules. Ontologies that define the meaning of semantic data play a key role in enabling the agent to understand what is on the Semantic Web, interact with sites and employ other automated services.

XPLANE

What Is the Killer App?



AFTER WE GIVE a presentation about the Semantic Web, we're often asked, "Okay, so what is the killer application of the Semantic Web?" The "killer app" of any technology, of course, is the application that brings a user to investigate the technology and start using it. The transistor radio was a killer app of transistors, and the cell phone is a killer app of wireless technology.

So what do we answer? "The Semantic Web *is* the killer app."

At this point we're likely to be told we're crazy, so we ask a question in turn: "Well, what's the killer app of the World Wide Web?" Now we're being stared at kind of fish-eyed, so we answer ourselves: "The Web is the killer app of the Internet. The Semantic Web is another killer app of that magnitude."

The point here is that the abilities of the Semantic Web are too general to be thought about in terms of solving one key problem or creating one essential gizmo. It will have uses we haven't dreamed of.

Nevertheless, we can foresee some disarming (if not actually killer) apps that will drive initial use. Online catalogs with semantic markup will benefit both buyers and sellers. Electronic commerce transactions will be easier for small businesses to set up securely with greater autonomy. And one final example: you make reservations for an extended trip abroad. The airlines, hotels, soccer stadiums and so on return confirmations with semantic markup. All the schedules load directly into your date book and all the expenses directly into your accounting program, no matter what semantics-enabled software you use. No more laborious cutting and pasting from e-mail. No need for all the businesses to supply the data in half a dozen different formats or to create and impose their own standard format.

"Cook" (sidestepping all the pages relating to cooks, cooking, the Cook Islands and so forth), find the ones that mention working for a company that's on your list of clients and follow links to Web pages of their children to track down if any are in school at the right place.

Agents

THE REAL POWER of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs. The effectiveness of such software agents will increase exponentially as more machine-readable Web content and automated services (including other agents) become available. The Semantic Web promotes this synergy: even agents that were not expressly designed to work together can transfer data among themselves when the data come with semantics.

An important facet of agents' functioning will be the exchange of "proofs" written in the Semantic Web's unifying language (the language that expresses logical inferences made using rules and information such as those specified by ontologies). For example, suppose Ms. Cook's contact information has been located by an online service, and to your great surprise it places her in Johannesburg. Naturally, you want to check this, so your computer asks the service for a proof of its answer, which it promptly provides by translating its internal reasoning into the Semantic Web's unifying language. An inference engine in your computer readily verifies that this Ms. Cook indeed matches the one you were seeking, and it can show you the relevant Web pages if you still have doubts. Although they are still far from plumbing the depths of the Semantic Web's potential, some programs can already exchange proofs in this way, using the current preliminary versions of the unifying language.

Another vital feature will be digital signatures, which are encrypted blocks of data that computers and agents can use to verify that the attached information has been provided by a specific trusted source. You want to be quite sure that a statement sent to your accounting program that you owe money to an online retailer is not a forgery generated by the computer-savvy teenager next door. Agents should be skeptical of assertions that they read on the Semantic Web until they have checked the sources of information. (We wish more *people* would learn to do this on the Web as it is!)

Many automated Web-based services already exist without semantics, but other programs such as agents have no way to locate one that will perform a specific function. This process, called service discovery, can happen only when there is a common language to describe a service in a way that lets other agents "understand" both the function offered and how to take advantage of it. Services and agents can advertise their function by, for example, depositing such descriptions in directories analogous to the Yellow Pages.

Some low-level service-discovery schemes are currently available, such as Microsoft's Universal Plug and Play, which focuses on connecting different types of devices, and Sun Microsystems's Jini, which aims to connect services. These initiatives, however, attack the problem at a structural or syntactic level and rely heavily on standardization of a predetermined set of functionality descriptions. Standardization can only go so far, because we can't anticipate all possible future needs.

The Semantic Web, in contrast, is more flexible. The consumer and producer agents can reach a shared understanding by exchanging ontologies, which provide the vocabulary needed for discussion. Agents can even "bootstrap" new reasoning capabilities when they discover new ontologies. Semantics also makes it easier to take advantage of a service that only partially matches a request.

A typical process will involve the creation of a "value chain" in which subassemblies of information are passed from one agent to another, each one "adding value," to construct the final product requested by the end user. Make no mistake:

to create complicated value chains automatically on demand, some agents will exploit artificial-intelligence technologies in addition to the Semantic Web. But the Semantic Web will provide the foundations and the framework to make such technologies more feasible.

Putting all these features together results in the abilities exhibited by Pete's and Lucy's agents in the scenario that opened this article. Their agents would have delegated the task in piecemeal fashion to other services and agents discovered through service advertisements. For example, they could have used a ***trusted*** service to take a list of ***providers*** and determine which of them are ***in-plan*** for a specified ***insurance plan*** and ***course of treatment***. The list of providers would have been supplied by another search service, et cetera. These activities formed chains in which a large amount of data distributed across the Web (and almost worthless in that form) was progressively reduced to the small amount of data of high value to Pete and Lucy—a plan of appointments to fit their schedules and other requirements.

In the next step, the Semantic Web will break out of the virtual realm and extend into our physical world. URIs can point to anything, including physical entities, which means we can use the RDF language to describe devices such as cell phones and TVs. Such devices can advertise their functionality—what they can do and how they are controlled—much like software agents. Being much more flexible than low-level schemes such as Universal Plug and Play, such a semantic approach opens up a world of exciting possibilities.

For instance, what today is called home automation requires careful configuration for appliances to work together. Semantic descriptions of device capabilities and functionality will let us achieve such automation with minimal human intervention. A trivial example occurs when Pete answers his phone and the stereo sound is turned down. Instead of having to program each specific appliance, he could program such a function once and for all to cover every ***local*** device that advertises having a ***volume control***—the TV, the DVD player and even the media players on the laptop that he brought home from work this one evening.

The first concrete steps have already been taken in this area, with work on developing a standard for describing functional capabilities of devices (such as screen sizes) and user preferences. Built on RDF, this standard is called Composite Capability/Preference Profile (CC/PP). Initially it will let cell phones and other nonstandard Web clients describe their characteristics so that Web content can be tailored for them on the fly. Later, when we add the full versatility of languages for handling ontologies and logic, devices could automatically seek out and employ services and other devices for added information or functionality. It is not hard to imagine your Web-enabled microwave oven consulting the frozen-food manufacturer's Web site for optimal cooking parameters.

Evolution of Knowledge

THE SEMANTIC WEB is not "merely" the tool for conducting individual tasks that we have discussed so far. In addition, if properly designed, the Semantic Web can assist the evolution of human knowledge as a whole.

Human endeavor is caught in an eternal tension between the effectiveness of small groups acting independently and the need to mesh with the wider community. A small group can innovate rapidly and efficiently, but this produces a subculture whose concepts are not understood by others. Coordinating actions across a large group, however, is painfully slow and takes an enormous amount of communication. The world works across the spectrum between these extremes, with a tendency to start small—from the personal idea—and move toward a wider understanding over time.

An essential process is the joining together of subcultures when a wider common language is needed. Often two groups independently develop very similar concepts, and describing the relation between them brings great benefits. Like a Finnish-English dictionary, or a weights-and-measures conversion table, the relations allow communication and collaboration even when the commonality of concept has not (yet) led to a commonality of terms.

The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web. This structure will open up the knowledge and workings of humankind to meaningful analysis by software agents, providing a new class of tools by which we can live, work and learn together. SA

Properly designed, the Semantic Web can assist the EVOLUTION of human knowledge as a whole.

MORE TO EXPLORE

Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. Tim Berners-Lee, with Mark Fischetti. Harper San Francisco, 1999.

An enhanced version of this article is on the Scientific American Web site, with additional material and links.

World Wide Web Consortium (W3C): **www.w3.org/**

W3C Semantic Web Activity: **www.w3.org/2001/sw/**

An introduction to ontologies: **www.SemanticWeb.org/knowmarkup.html**

Simple HTML Ontology Extensions Frequently Asked Questions (SHOE FAQ): **www.cs.umd.edu/projects/plus/SHOE/faq.html**

DARPA Agent Markup Language (DAML) home page: **www.daml.org/**

Semantic Focus
All about the Semantic Web

Subscribe

Looking for something?

Search

SEP 24th 2007

Introduction to the Semantic Web Vision and Technologies - Part 1 - Overview

Published **5 years ago** by **Cody Burleson**

Tweet 2

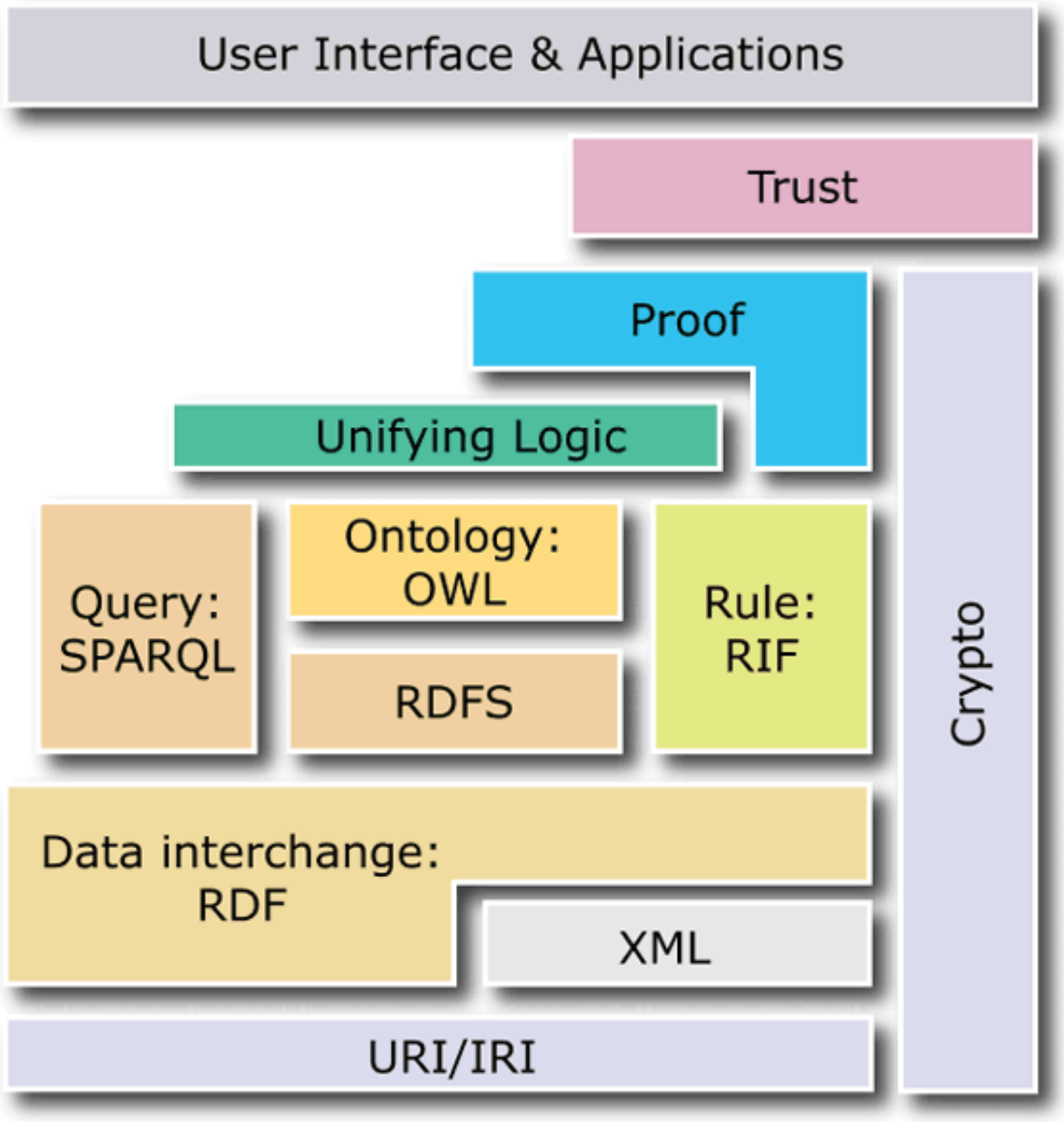
The World Wide Web has long been evolving towards the vision of the Semantic Web — an extension of the existing web through which machines are better able to interoperate and work on our behalf. It promises to infuse the Internet with a combination of metadata, structure, and various technologies so that machines can derive meaning from information, make more intelligent choices, and complete tasks with reduced human intervention. It is a dramatic vision that stands to transform the existing Web in devastatingly powerful ways.

It is also a realistic vision. In some ways, in fact, it is already here. Semantic Web standards and technologies are maturing, several tools exist, and new applications are frequently emerging. Similar to the early days of the existing Web, the vision awaits only understanding, acceptance, and perhaps a few "killer apps" that will deliver on its promise and prove its transformational value to the world.

This is the first of a series of articles written exclusively to help the Semantic Focus community understand the Semantic Web vision and technologies. If you are new to Semantic Web concepts then you might have at least learned how unforgiving and overly academic the existing material can be. At least, that is what I have always thought, but I may just be a dunce.

While I do not claim to be a guru, I know that I can help those of you who may be taking your first steps. I will deliver a logical progression of concepts designed to get you up to speed as quickly and painlessly as possible. I

hope to deliver one part of this series per week using a version of the Semantic Web technology stack as a framework. The stack (aka *the Semantic Web layer cake*) is a rather famous illustration of the key Semantic Web enabling technologies. Building one upon another from bottom to top, these technologies can help us realize the full Semantic Web vision. To my knowledge, all of these exist in various forms of maturity and you can now use one, some, or all of them to empower your ideas.



In this series, we will work our way up from the bottom of the stack, eating one layer of the cake at a time. Along the way, we'll take a few pleasant detours to indulge in the ridiculous joy of programming and also try out some free tools. When the series is complete, you should be able to:

- Understand and articulate the Semantic Web vision.
- Know the basic framework and core technologies.
- Be aware of some available tools.
- Understand how mature the technology is and how you might (or might not) use it today.
- Get involved for further learning, contribution, or just plain geek fun.

So, let's begin at the beginning — the part where everyone at some point or another asks, "What is the Semantic Web?"

The Semantic Web is...

> "The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."
>
> - *Tim Berners-Lee, James Hendler and Ora Lassila; Scientific American, May 2001*

 The Semantic Web is a vision for the future of the Web, a "Web of meaning" (i.e. *semantics*), that was set forth by Tim Berners-Lee. Sir Berners-Lee is often referred to as the father of the Web. He's sort of like that white-bearded old man in Matrix Reloaded (the one who says 'ergo' a lot — the Architect) only not as pompous: "I just had to take the hypertext idea and connect it to the TCP and DNS ideas and — ta-da!" He's the man who originally wanted to establish this single global space to create the World Wide Web and that's how HTML and HTTP came into existence. He went on to establish the World Wide Web Consortium (W3C) and then later thought of embedding large amounts of machine-understandable metadata into documents; this is what gave rise to the Semantic Web.

You see, most of the Web's content is designed for humans to read and it is not very easy for computer programs to manipulate that information meaningfully. Though the information is semi-structured, the structures typically only defines how content should be rendered in a client browser. HTML does not inherently tell us anything about the subject and nature of the content. So, the idea is to come up with some standard ways to express better meaning around information so that computers can help us use the information more effectively. This simple idea is the very essence of the Semantic Web.

Imagine a scenario, for example, where software agents can roam the web and carry out sophisticated tasks on our behalf. This is different than searching content for keywords and popularity. It is a web where computers are able to infer meaning from content based on metadata and assertions that have already been made. It is a web where information can be automatically classified and related through the help of reasoning engines and description logics. Or, in a more practical sense, it's a web where services can be found, integrated, and invoked more easily and more dynamically.

Some like to call this scenario "Web 3.0." Others have called it the "dark web" (where computers are using the Web more than we are; churning through information for our foreground benefit). Some people, of course, do not agree with it at all (believing it to be overly ambitious). Others interpret it on sci-fi tangents (i.e. a one-brain global intelligence and what-not).

Personally, I just call it the Semantic Web — a good vision, a practical blueprint, and a set of tools and technologies can be incredibly useful once you come to understand them. If you wanted to summarize it in a slide presentation for you manager, you might do it like this:

The Semantic Web...

- Is a vision for the future Web (a web of meaning — semantics); originally defined by Tim Berners-Lee (aka father of the Web).
- Is not a separate web, but an extension of the current one.
- Provides a way for machines to get much better at being able to process and understand the data that they merely display at present.
- Is a web on which machine reasoning can become ubiquitous and powerful.
- Describes an emerging set of standards, markup languages, and related processing tools.
- Is a rolling snowball; interest and momentum is building fast — heads up!

If you want to be persuasive, you might also add that according to the 2006 Semantic Conference, semantics is already a 2 billion per year market and is projected to grow to over 50 billion by the year 2010. That's 20 billion more per year than the market for pizza — yum!

The first steps of weaving the Semantic Web into the structure of the existing web are indeed already under way. The foundation has been laid; the rest is up to us. To realize the vision, however, we must begin understanding it on a much deeper level and we've got to roll up our sleeves to get our hands dirty. That is exactly what we'll start doing next week when we dive into the layer-cake to munch on the basic enabling technologies such as Unicode, URI, and XML. Until then, enjoy your work and the Web; it is an exciting time to be alive!

Tags:

- Semantic Web Introduction

About the author

Cody Burleson is President and CEO of Burleson Technology Group, LLC. BTG provides Information Workplace and business integration solutions on leading enterprise platforms and is currently developing a semantic Workplace application. ~~BTG also sponsors Workplace Design, a wiki community for sharing news, information, tools, and notes on the Information Workplace and business integration.~~

If you liked this entry, you might like the following:

- 30+ Semantic Web Introductions, References, Guides, and Tutorials
- Introduction to the Semantic Web Vision and Technologies - Part 2 - Foundations
- 17 Semantic Web, RDF, and OWL Videos

The Data Administration Newsletter
Since 1997

Data Modeling, RDF, & OWL - Part One: An Introduction To Ontologies

by David C. Hay
Published: April 1, 2006

(Article URL: http://www.tdan.com/view-articles/5025)

Published in TDAN.com April 2006

[*This is the first of three articles discussing the new/old ideas of semantics and ontology and how they affect the way we analyze data. This article introduces the main concepts, and the second article will show an example of converting a data model to the web ontology language, OWL.*]

Everyone knows that we are drowning in information, both from the databases in our companies as well as from the world-wide web, the media, and life in general. The information technology industry has been wrestling with this problem for years, and one is entitled to wonder if things will ever get better.

Well, there are a couple of new/old ideas on the horizon that might help: semantics and ontology.

Data modeling was invented three decades ago to assist in the design of databases-in particular relational databases. As it matured, the technique has become recognized as a tool for analyzing the semantics of an organization-what is the structure of the organization's information as it is used in carrying out its mission?

In recent years, from a completely different direction, the artificial intelligence world, semantics has arisen as a subject of interest in its own right. This came the artificial intelligence world's desire to create computerized natural language processors.

These two fields are finally coming together, and this article is an attempt to articulate that link.

In particular, companies are beginning to recognize that semantics is important if their systems (and their people, for that matter) are going to communicate with each other, and, based on this recognition, they are also recognizing the importance of collecting "ontologies", or glossaries that describe the language they use to carry out their activities.

In other words, a couple of 2500 year-old words are becoming the hot new buzzwords in our industry.

Semantics is the Greek philosophic study of the nature of meaning, especially as it is expressed in language. It is the "study of the signification of signs and symbols, as opposed to their formal relations (syntactics)."[1]
Ontology is another branch of Greek philosophy, "concerned with identifying, in the most general terms, the kinds of things that actually exist."[2]

In other words, ***ontology*** tells us *what exists*. ***Semantics*** tells us *how to describe it*.

WHAT IS A DATA MODEL?

A ***data model*** is a drawing that represents data "things" and relationships between them. The meaning of the model varies, depending on its purpose:

- It can represent a *data base design*, with the boxes representing tables and the lines representing foreign keys. Also represented are the columns of the tables.

- It can be a ***conceptual*** model representing the structure of a business, with the boxes representing things of significance to the business and the lines representing semantic relationships between them. Also represented may be the definitions of data describing those things of significance.

These are very different things. A business (or "conceptual") data model captures the semantics of an organization for the purpose of both communicating both with the business community and providing and architecture for database and system design. A database design describes an artifact that can be employed to store and manipulate data.

Other than constraints on cardinality, *business rules* are not generally represented on data models of either kind. Even in the case of business data models, the models are supposed to represent fundamental structures, while business rules represent variable constraints.

In other words, database design, business data modeling, and business rule modeling are three very different things. They do, however, represent a particular mindset, which for purposes of this article, we will characterize simply as the "data modeling mindset".

This article then uses that to describe a completely different mindset.

About Data Models and Ontology Languages

A conceptual data model is, of course, a kind of ontology. It is about defining *categories* of data. Its graphic nature provides an excellent basis for discussing and negotiating the meaning of those categories. Accompanied by business rules analysis, the two provide a basis for collecting data according to those categories, and its corresponding database design provides a mechanism for doing so. The point is that data models are to be understood by humans, with computers only serving as gateways to permit capture of "valid" data.

In its latest incarnations, however, an ontology language begins with *instances* of actual data. It's purpose is to classify them so that computers can make inferences from them.

The data modeling mindset is based upon the ***closed world assumption***:

> *Only that which is asserted is known.*

Ontology languages are based on the ***open world assumption***.

> *All assertions are assumed to be true until proven otherwise.*

This means that when you build a system using a data modeling approach:

- You can only enter data that you know to be valid.
- You are "encouraged" to enter complete information.
- There are no other data.
- The data model entity classes and their derived tables are templates.

With an ontology database:

- You can enter what you know to be true.
- You can enter incomplete information.
- You (and the computer) can *infer* other things.
- Ontology classes are simply sets of things.

This is a profoundly different view of the world, as we shall see, below.

About the Semantic Web

Before going into ontology languages in detail, it is worth taking a moment to understand "The Semantic Web". As imagined by Tim Berners-Lee, the inventor of the World-wide Web, the "first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web-a web of data that can be processed directly or indirectly by machines".[3]

Ok, so how is that different from simply creating a networked database? If all validation of data is in a program, the program acts as a filter, the way we discussed before. If, on the other hand, data are stored with the semantics visible to a wide range of processors, then the data are more powerful, and the opportunities for discovering new things in them is greater. Michael Daconta and his colleagues describe four stages in the smart data continuum:

- ***Text and databases (pre-XML)***-Most data are proprietary to an application. The "smarts" are in the application and not in the data.
- ***XML documents in a single domain***-Here data achieve application independence within a domain. For example XML could describe standard semantics within the health care industry, the insurance industry, and so forth.
- ***Taxonomies and documents with mixed vocabularies***-In this stage data can be collected from multiple domains and accurately classified. This classification can then be used for discovery of data. Simple relationships between categories in the taxonomy can be used to relate and thus combine data. Data are now smart enough to be easily discovered and sensibly combined with other data.
- ***Ontologies and rules***-in this stage, new data can e inferred from existing data by following logical rules. Data are not smart enough to be described with concrete relationships and sophisticated formalisms. Logical calculations can be made on this "semantic algebra". In this stage data no longer exist as a blob

but as a part of a sophisticated microcosm.[4]

The semantic web, then, is an extension of the World-wide web to allow for not just the retrieval of documents based on key words, but for their retrieval based on the semantics of their contents.

The semantic web is based on the idea of a "layered architecture". Much like the ISO concept of layers in data communications, the semantic web architecture is composed of the following layers:

- ***URIs and Namespaces***-the names of things
- ***XML and XMLS Data types***-a means of communicating data
- ***RDF and RDF/XML***-a basic language
- ***RDF Schema and Individuals***-an ontological primitive
- ***Ontology languages, such as OWL***-the logical layer
- ***Applications***-the implementation layer.

The field is new, and it is not clear to this author just what the "Applications" layer might look like. But we can address the others. Specifically, RDF and OWL represent structured languages for representing ontologies that we can map back to what we are used to doing with data models.

Uniform Resource stuff

In order to talk about something, it is necessary to name it. The semantic web provides a scheme for naming things in two layers.

First of all, the general concept of a *Uniform Resource Identifier* (URI) is simply a formatted identifier that identifies anything. The name is in two parts:

- A ***scheme name***, and
- A ***scheme-specific name.***

There is no outside control over URIs, so they can be whatever you want them to be, such as:

hay:david

Note, of course, that within the context of a particular ontology, all URIs must be unique.

A *Uniform Resource Locator* (URL) is a URI that is specifically used to locate resources on the World-wide Web. The scheme name and the first elements of the scheme-specific name are regulated to insure uniqueness across the World-wide Web. To call a particular URL your own, you have to get permission from the Internet Corporation for Assigned Names and Numbers (ICANN). For example, the following is the URL of your author's company:

http://www.essentialstrategies.com

The literature also describes a *Uniform Resource Name*, but the descriptions are contradictory and confusing. It apparently comes down to a URI whose scheme is "urn:".

Namespaces

An *XML namespace* is the URI that describes an ontology from which terms are taken. As you will see, in this context XML is the language that is used to describe an ontology. Since the description of an XML namespace can be lengthy, a prefix is usually assigned to each, in order to simplify referring to a term.

For example, the set of terms that define the OWL language is itself an ontology, defined in XML. Its namespace, then is described as follows:

```
xmlns:owl="http://www.w3.org/2002/07/owl"
```

In the OWL namespace, then, the term "class" would be described thus:

```
xmlns:owl="http//www.w3.org/2002/07/owl#class"
```

Once the namespace is declared, however, this can be abbreviated:

```
owl:class
```

XML and XML Schema

As mentioned above, the RDF and OWL languages are expressed in XML.

Here's a whirlwind synopsis of XML:

An XML document contains "tags" describing strings of text. These are similar to the tags in HTML, but where HTML tags describe formatting components of a document, these tags describe the semantic content of it. For example,

```
<product>
  <product name>BlackBerry</product name>
</product>
```

As you can see, the tag describes the text that follows. The text is then demarked by a corresponding end tag in the form .

Tags are typically defined in accompanying files called *data type definitions* (DTD). A DTD is itself a document with the following structure:

- ***Header:*** <DOCTYPE PRODUCT>
- ***Context for the tag:*** <!ENTITY **product** (**product_name**)>
- ***Tag definition:*** <**product_name** (#PCDATA)>

Note that "(#PCDATA)" simply means that's where actual data go. In the context line, a character may be added after the tag name (for example product_name+). The character determines how many occurrences of the tag are required for each occurrence of the context tag:

- ***(no character)*** - (Default) mandatory single valued (must be ... one and only one.)
- \+ - Mandatory one or more occurrences (must be ... one or more).
- ? - optional, single valued (may be one and only one).
- * - optional, one or more occurrences (may be one or more).

XML schema is an alternative to DTDs. XML Schema is an XML document that configures other documents.

RDF and OWL are defined as tags in XML Schemas. An ontology is defined as a namespace, and terms are described as elements of that namespace. For example, the Ontology "contact" might be used as follows (note that RDF itself must be defined first):

```
<?xml version="1.0"?>

    <rdf:RDF
       xmlns:rdf=
          http://www.w3.org/1999/02/22-rdf-syntax ns#

        xmlns:contact=
          "http://www.w3.org/2000/10/swap/pim/contact#">

        <contact:person
          rdf:about="http://www.w3.org/People/EM/contact#me">

        <contact:fullName>David Hay</contact:fullName>

        <contact:mailbox
            rdf:resource="mailto:tdan@davehay.com"/>

        <contact:personalTitle>Mr.</contact:personalTitle>

        </contact:Person>

    </rdf:RDF>
```

First the contact namespace is defined with the name "contact", and the terms "person", "fullName", "mailbox" and "personalTitle" are used to capture values. The paragraph above asserts that person (me) is described by a full name "David Hay", my mailbox is "tdan@davehay.com", and my (personal) title is "Mr."

Note that contact:person is equivalent to:

http://www.w3.org/2000/10/swap/pim/contact#person.

RDF and RDF/XML

Resource Description Framework is the basic language layer for data representation. It is rendered in XML, and consists of a preliminary set of tags for describing semantics. RDF can be used as metadata to describe documents and images. Its most important tags are:

```
<rdf:subject>

<rdf:predicate>

<rdf:object>
```

RDF and Data Modeling:

RDF tags that correspond to data modeling constructs are the following:

- ***Entity class:***: <rdf:subject>
- ***Relationship:*** <rdf:predicate> + <rdf:object>
- ***Attribute:*** <rdf:predicate>an attribute of</rdf:predicate>

For example,

<rdf:subject>Kleenex brand tissues</rdf:subject>

<rdf:predicate>are sold by</rdf:predicate>

<rdf:object>

Kimberly Clark Corporation

</rdf:object>

Note that there is no distinction between classes and instances.

RDF Schema

Resource Description Framework Schema (RDFS) is an extension of RDF. In addition to the RDF tags available are additional tags to define:

- Resource
- Class
- Sub-class
- Range
- Domain

and others.

Interestingly enough, some RDF tags are actually defined using RDFS tags.

RDFS and Data Modeling

RDFS tags that correspond to data modeling constructs are the following:

- ***Entity class:*** <rdfs:class>
- ***Attribute:*** <rdfs:domain>
- ***Relationship:*** <rdfs:range>
- ***Instance:*** <rdf:type>

In RDFS, attributes and relationships are properties that are defined *before* assigning them to classes.

For example, imagine an Essential Strategies, Inc. ontology called "MRP", containing the concept "manufactured by". This could be defined as a property as follows:

```
<rdf:property
   rdf:about="http://www.essentialstrategies.com/MRP
   #manufactured by">

     <rdfs:domain>Product</rdfs:domain>

     <rdfs:range>Party</rdfs:range>

  </rdfs:property>
```

Here, the relationship "manufactured by" being defined in terms of the classes it relates. Contrary to the way data modelers use the word "domain", here ***domain*** is the class that is on the first end of the relationship. ***Range*** refers to the class that is on the second end of the relationship. A relationship is considered a property of the first class.

Note that all relationships and attributes are considered optional many-to-many. There are no cardinality constraints in RDF.

Ontology Languages

As you can see, even RDFS is quite limited in its ability to express semantic constructs. Most notably, it doesn't allow expression of constraints. Moreover, it has few descriptors to make extensive inferences. Thus it is not really expressive enough to support the Semantic Web.

As part of its IDEF series of notations, the Federal Government has sponsored the creation of IDEF5, an ontology expression graphical language. A report describing it was published in 1994[5], but it has gotten little publicity since then. The report is an excellent overview of ontological topics, and the approach is very thorough.

Since IDEF5, like data modeling, is a graphical approach to ontological modeling, however, it does not serve the purposes of the Semantic Web. The ***Web Ontology Language*** (OWL)* was developed to provide a syntax that can be understood directly by computers. OWL builds on RDF and RDFS, and like them, it is constructed from XML tags.

There are actually three versions of OWL: OWL Lite, OWL DL, and OWL Full. The nuances of the differences among these are beyond the scope of this paper, so we will focus on OWL DL.

Remembering the differences we described above between data modeling and ontology languages, it is important to reiterate that the structures to be built using OWL are not restrictive. The open world assertion applies. Beginning with a series of instances and a series of assertions, the assumption is that anything can be true unless asserted otherwise.

(Also, accept the fact that OWL is primarily intended to be understood by computers, not people. Hence the discussions which follow will seem pretty arcane to the graphically-motivated among us.)

For example, in an airports database, you might have something called "AA243". Absent any other information, that could refer to either a flight or an airport. After you have asserted that it is in fact a flight number, you have *not ruled out* that it might also refer to an airport. To prevent that, you must explicitly declare that the class of airports and the class of flights is ***disjoint.*** That is, the same thing cannot be an instance of both classes.

This open assumption is significant, because, given a large amount of data from disparate systems, it is possible that a computerized analysis along these lines might show up things that people would never figure out. To be sure, there will be a lot of nonsense assertions initially, until people learn to clarify the rules, but even that exercise will be useful in helping people better understand their data. This allows data from systems where the language is not quite consistent to at least be viewed collectively, and, with luck, those inconsistencies themselves will become clear.

The idea is to begin with instances and classify them by their properties. Your author's son showed an early propensity for philosophy when, at the age of three, he decided to build a collection of red things. First he sorted his toy trucks, his action figures, and his other toys to gather together the red ones. Then he went around the house to find his mother's lipstick, one of his father's shirts, and various other (red) paraphernalia.

This is what OWL does. Among other things, it classifies things by their properties.

OWL and Data Modeling

The OWL tags that correspond to data modeling constructs include the following:

- ***Entity Class:*** <owl:**class** rdf:id=". . .">
- ***Attribute:*** <owl:**datatypeProperty** rdf:id=". . .">
- ***Relationship:*** <owl:**objectProperty** rdf:ID=". . .">

Note that, as with RDF, both attributes and relationships are ***properties*** that must be defined first before being attached to classes.

For example, here are two classes:

```
<owl:class rdf:ID="Class1">

<owl:class rdf:ID="Class3">
```

The attribute "Attr5" is in fact an attribute of both of these classes:

```
<owl:DatatypeProperty rdf:ID="ATTR5">
   <rdf:Type rdf:resource="owl:#FunctionalProperty"/>
   <rdfs:domain rdf:resource="#Class1"/>
   <rdfs:domain rdf:resource="#Class3"/>
</owl:DatatypeProperty>
```

That the DatatypeProperty is a "functionalProperty" means that the property can have no more than one value in each domain. It is not required to have any values. This is appropriate for attributes in a relational environment, which are not allowed to have more than one value. For relationships, this is useful, although since it does not require a value, it is not adequate for the full range of cardinality issues. There are other problems with it in relationships, which we will discuss below.

RDF had a tag "type" that was supposed to allow you to specify instances of classes, but since it didn't have classes, that didn't make much sense. Now that OWL has classes, we can talk about instances of classes. For example:

```
<owl:Class rdf:ID="Ship"/>

<owl:Thing rdf:ID="Titanic">

    <rdf:type rdf:resource="#Ship">

</owl:Thing>
```

There are two approaches to specifying relationships in OWL.

In the first case, an ObjectProperty is simply defined, where the domain and range are part of the definition. For example:

```
<owl:Class rdf:ID="City"/>

<owl:Class rdf:ID="State"/>

<owl:ObjectProperty rdf:ID="located_in_state">
    <rdfs:domain rdf:resource="#City"/>
    <rdfs:range rdf:resource="#State"/>

</owl:ObjectProperty>
```

Thus since the domain and range are part of the definition of the object property, the name must be unique to this entity class pair. You cannot assign this relationship to any other class pair.

An alternative is to define the property without specifying a domain and range. In this case, you then define a class as being a sub-class of a restriction that applies the property. In this case many class pairs can use the same object property. For example:

```
<owl:ObjectProperty rdf:ID="located_In"/>

<owl:Class rdf:ID="City">
  <rdfs:subClassOf>

      <owl:restriction>

        <owl:onProperty rdf:resouce="located_In"/>

        <owl:someValuesFrom rdf:resource="State"/>

      </owl:restriction>
  </rdfs:subClassOf>

</owl:Class>
```

So, it is possible to convert an ontology represented by a data model into one represented by an ontology language. The model assumes constraints we don't normally realized (like disjointedness), and it will be important to introduce any business rules we've identified as well.

Summary

Data modeling, database design, and business rule modeling are all part of a particular way of looking at the world. The semantic web and the ontology languages that support it are part of a new way of looking at the world. The differences are in terms of premises, the way classes are identified, and the implications of constraints.

- ***Premises***
    - Data modeling, etc.
    - "Closed" world
    - Only what is asserted is true
    - Ontology languages
    - "Open" world
    - Anything may be true if it does not conflict with assertions
- ***Approach to classes***
    - Data modeling, etc.
    - Begin with class definitions of fundamental categories
    - Define attributes
    - Identify sample instances
    - Ontology languages
    - Begin with instances
    - Identify attributes
    - Define classes based on attributes
- ***Constraints and business rules***
    - Data modeling, etc.
    - Determine what data are acceptable
    - Reject data that do not conform
    - Ontology languages
    - Assert what is known to be true
    - Infer what else may be true.

As an example, consider the typical data modeling assertion:

Each CITY must be located in *one and only one STATE*.

To a data modeler, this implies the following:

1. If "Portland" is entered as a CITY without a STATE identified, it is not acceptable.
2. If "Portland" is entered as a CITY and located in STATE "Maine", then a record with the CITY "Portland" located in state "Oregon" is not accepted.

To an ontologist, however, this implies the following:

1. "Portland" may be entered without specifying the STATE.
2. If "Portland" is entered "located in" "Maine", and "Portland" is identified as a CITY then "Maine" must be a STATE.
3. If, in addition to statement 2., "Portland" is entered as "located in" "Oregon", then "Oregon" must be a STATE, and either:
    - "Oregon" and "Maine" must be two names referring to the same STATE, or
    - The CITY referred to by the name "Portland" in "Oregon" must be a *different* CITY than the one referred to by the name "Portland" in "Maine".

Interesting, yes? For an example of converting a data model to OWL, tune in next quarter for the next article on this subject.

---

[1] G. Kemmerling, *Philosophical Dictionary*, 2002.

[2] Ibid., http://www.philosophypages.com/dy/o.htm#onty.

[3] Tim Berners-Lee, *Weaving the Web*. Harper, San Francisco. 1999.

[4] Michael C. Daconta, Leo J. Obrst, Keven T. Smith, *The Semantic Web*. Wiley, Indianapolis. 2003.

[5] Knowledge Based Systems, Inc. *IDEF5 Method Report*. Prepared for Armstrong Laboratory AL/HRGA Wright-Patterson Air Force Base, Ohio. Can be found at http://www.idef.com/pdf/Idef5.pdf.

* You may wonder why the "Web Ontology Language" has the acronym "OWL". It seems that In *Winnie the Pooh*, Owl imagines that his name is spelled "WOL", until his friends correct him. Here, the World Wide Web Consortium (W3C) decided to start with the correct spelling.

Go to Current Issue | Go to Issue Archive

Recent articles by David C. Hay

- From Data Modeling to Ontologies: Discovering What Exists, Part 2

- From Data Modeling to Ontologies: Discovering What Exists, Part 1
- Book Review: Principles of Data Management:

David C. Hay - In the information industry since it was called "data processing," Dave Hay has been producing data models to support strategic and requirements planning for more than twenty-five years. As President of Essential Strategies, Inc. for nearly twenty of those years, Dave has worked in a variety of industries including, among others, banking, clinical pharmaceutical research, broadcasting, and all aspects of oil production and processing. Projects entailed various aspects of defining corporate information architecture, identifying requirements, and planning strategies for the implementation of new systems.

Dave's recently published book, *Enterprise Model Patterns: Describing the World*, is an "upper ontology" consisting of a comprehensive model of any enterprise from several levels of abstraction. It is the successor to his groundbreaking 1995 book, *Data Model Patterns: Conventions of Thought* – the original book describing standard data model configurations for standard business situations.

In between, he has written *Requirements Analysis: From Business Views to Architecture* (2003) and *Data Model Patterns: A Metadata Map* (2006). Since he took the unusual step of using UML in the *Enterprise Model Patterns…* book, a follow-on book, *UML and Data Modeling: A Reconciliation* was published later in 2011. This book both shows data modelers how to adapt the UML notation to their purposes, and UML modelers how to adapt UML to produce business-oriented architectural models.

Dave has spoken at numerous international and local DAMA, semantics, and other conferences as well as at various user group meetings. He can be reached at dch@essentialstrategies.com, (713) 464-8316, or via his company's website.

Quality Content for Data Management Professionals Since 1997

Scalable Interoperability Through the Use of COIN Lightweight Ontology

Hongwei Zhu[1,2] and Stuart E. Madnick[1]

[1] Massachusetts Institute of Technology
Sloan School of Management
30 Wadsworth Street, E53-320, Cambridge, MA 02142, USA
{mrzhu,smadnick}@mit.edu
[2] Old Dominion University
College of Business and Public Administration
Constant 2079, Norfolk, VA 23529, USA
hzhu@odu.edu

Abstract. There are many different kinds of ontologies used for different purposes in modern computing. A continuum exists from lightweight ontologies to formal ontologies. In this paper we compare and contrast the lightweight ontology and the formal ontology approaches to data interoperability. Both approaches have strengths and weaknesses, but they both lack scalability because of the n^2 problem. We present an approach that combines their strengths and avoids their weaknesses. In this approach, the ontology includes only high level concepts; subtle differences in the interpretation of the concepts are captured as context descriptions outside the ontology. The resulting ontology is simple, thus it is easy to create. It also provides a structure for context descriptions. The structure can be exploited to facilitate automatic composition of context mappings. This mechanism leads to a scalable solution to semantic interoperability among disparate data sources and contexts.

Keywords: lightweight ontology, formal ontology, context, mediation, scalability, semantic heterogeneity.

1 Introduction

Ontologies have been widely used in modern computing for purposes such as communication, computational inference, and knowledge organization and reuse [7]. For different purposes there are a variety of different ontologies, which range from a glossary, to a taxonomy, a database schema, or a full-fledged logic theory that consists of concepts, relationships, constraints, axioms, and inference machinery. As illustrated in [21], a variety of ontologies form a continuum from lightweight, rather informal, to heavyweight, and formal ontologies.

The lightweight ontology approach and the formal ontology approach are often used differently and have different strengths and weaknesses. Both approaches can be used to support data interoperability among disparate sources.

M. Collard (Ed.): ODBIS 2005/2006, LNCS 4623, pp. 37–50, 2007.

Lightweight ontologies usually are taxonomies, which consist of a set of concepts (i.e., terms, or semantic types) and hierarchical relationships among the concepts. As an artifact, it is relatively easy to construct a lightweight ontology. However, such lightweight ontologies do not capture the detailed semantics of the concepts, which sometimes is documented in a data dictionary, and/or embedded in the data models and the data processing programs.

There are two different approaches to using lightweight ontologies for interoperability purposes. One approach is to develop a single lightweight ontology, in which case all parties need to agree on the exact meaning of the concepts. The lightweight ontology and the agreements together form a standard that all parties uniformly adopt and implement. That is, a lightweight ontology is often used to support strict data standardization. However, reaching such agreements can be difficult. For example, a data standardization effort within the U.S. Department of Defense (DoD) took more than a decade only to standardize less than 2% of the data across all organizations of the DoD [18]. The alternative approach is to allow multiple lightweight ontologies to co-exist, in which case mappings among the ontologies need to be provided. Because the semantics is not formally captured in the ontologies, efforts are required to identify the semantic differences and then develop (often hand-code) the mappings to enable pair-wise interoperability. The number of pair-wise mappings is $n(n\text{-}1)$ (which is $O(n^2)$) if there are n different ontologies, thus the amount of effort required increases quickly as n becomes large. This is the so called n^2 problem of data interoperability. A survey [19] shows that approximately 70% of the costs of data interoperability projects are spent on identifying the semantic differences and developing code to reconcile them.

In contrast, the formal ontology approach uses axioms to explicitly represent semantics and has inference capabilities. This approach can also support interoperability either via a single ontology or via mappings of multiple ontologies. The key difference is that the semantics of the ontological concepts and the mappings are explicitly captured in a formal logic theory.

To summarize, both ontology approaches can be used to support data interoperability either via standardization or via mappings of multiple ontologies. The difficulty of reaching an agreement on a single data standard can be enormous so that in practice multiple standards (i.e., ontologies) co-exist even within a single organization. Thus, in practice ontology mappings are required to enable interoperability among data sources and systems. Both ontology approaches suffer from the n^2 problem. The key difference between the two ontology approaches is that lightweight ontologies do not capture the semantics in the ontologies, whereas formal ontologies explicitly capture semantics. As artifacts, lightweight ontologies are simple and easy to create, whereas formal ontologies are complex and difficult to create. But the semantics and the mappings of lightweight ontologies are often scattered in various data models and data processing programs, making maintenance extremely difficult. The semantics and mappings of formal ontologies are in the form of a logic theory, which is relatively easier to maintain. Both approaches have weaknesses that limit their effectiveness.

It is desirable to have an approach that combines the strengths and avoid the weaknesses of the two ontological approaches. In this paper, we present such an approach, which is developed in the COntext INterchange (COIN) project [3, 5, 25]

for semantic data interoperation purposes. It uses a lightweight ontology, which provides the structure for organizing context descriptions to account for the subtleties of the concepts in the ontology. We will use the terms *COIN ontology* and *COIN lightweight ontology* interchangeably. COIN also implements a reasoning algorithm to determine and reconcile semantic differences between different data sources and receivers.

The rest of the paper is organized as follows. In Section 2, we describe the COIN lightweight ontology approach. In Section 3, we present the scalability benefit of the approach. In Section 4, we discuss related work. In Section 5, we conclude and point out future research.

2 COIN Lightweight Ontology

We will use an online price comparison example to illustrate the COIN lightweight ontology approach.

2.1 Online Price Comparison Example

Numerous vendors make their pricing information available online. With web wrappers, such as Cameleon [2] and others [1], and the increasing adoption of XML and web services, one can gather price data and compare offers from different vendors. To perform meaningful comparisons, one has to reconcile the semantic differences of price data, especially when data is from vendors scattered around the world [22].

Consider a scenario where data is from 30 vendors from 10 different countries. For simplicity of discussion in this paper, let us assume that all vendors quote prices using the same schema and same *Product* identification, represented using the following first order predicate:

quote(*Product, Price, Date*)

but different vendors use different conventions so that the price values are interpreted differently depending on which vendor provides the quote. Table 1 provides a few examples of different interpretations of price. A *base price* refers to price with taxes and shipping & handling (S&H) excluded (e.g., price quotes from vendors 2 and 3).

Let us assume that each vendor uses a different convention, thus we have 30 unique conventions, which we call *contexts*. We can label vendor *i*'s context as c_i. For

Table 1. Interpretations of Price

Vendor	Interpretation of Price
1	In 1's of USD, taxes and S&H included
2	In 1's of USD, taxes and S&H excluded
3	In thousands of Korean won, taxes and S&H excluded
…	…
30	In millions of Turkish lira, taxes included

simplicity, we will assume that users normally adopt a vendor context. Or we can assume that the only users are the vendors, each of whom wants to compare his prices with all of his world-wide competitors and wants the comparison done in his own context. In this scenario, to allow users in all contexts to meaningfully compare vendor prices, it is necessary that price data from other contexts be converted to the user context, which would require 870 (i.e., 30*29=870) conversions. Hand-coding these conversions and maintaining them over time, since contexts do change (e.g., prices in French francs and German deutschemarks became Euros), can be costly and error-prone.

2.2 COIN Lightweight Ontology

In the example, there are a number of subtle differences in the meaning of the high level concept *price*. It is important that these subtleties are captured and the differences are reconciled for meaningful comparisons.

Like the traditional lightweight ontology, the COIN ontology includes a set of concepts, among which there can be a hierarchy represented with an *is_a* relationship. Besides, the COIN ontology also includes *attribute* as a binary relationship between a pair of concepts. Attributes are also called roles, and correspondingly attribute names are called role names. For example, *price* can be the *hasPrice* attribute of *product*. Conversely, *product* can be the *priceOf* attribute of *price*. To capture the subtle differences in meaning, the COIN lightweight ontology introduces *modifier* as a special kind of attribute. The values of modifiers are specified as context descriptions outside the ontology. Fig. 1 shows a graphic representation of the COIN lightweight ontology for the online price comparison example.

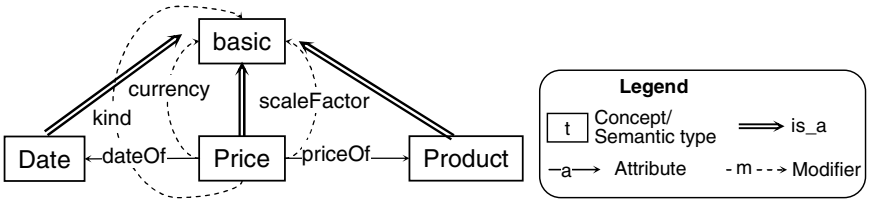


Fig. 1. COIN lightweight ontology for online price comparison example. It contains only high level concepts, the refined variants of which can be derived from the assignments of modifiers that belong to each high level concept.

In this ontology, we include a modifier-free root concept *basic*, which is similar to *thing* as the root in many object-oriented models. We include three modifiers: *kind*, *currency*, and *scaleFactor*. Each modifier captures a particular aspect in which the underlying concept can have different interpretations. Contexts are described by assigning values to modifiers present in the ontology. In simple cases, a specific value is assigned to a modifier in a context. In other cases, the assignment must be specified by a set of rules. In either case, a context is conceptually a set of assignments of all modifiers and can be described by a set of <*modifier*, *value*> pairs. For example, contexts c_2 and c_3 (refer to vendors 2 and 3 in Table 1) can be described as:

c_2 := { <*kind*, *basePrice*>, <*currency*, *usd*>, <*scaleFactor*, 1> }

c_3 := { <*kind*, *basePrice*>, <*currency*, *krw*>, <*scaleFactor*, 1000> }

The language used in COIN for describing context (as well as context mappings and the lightweight ontology) is based on F-logic [12], an object-oriented logic. F-logic rules are converted to Datalog for reasoning purposes. In COIN, various "user-friendly" front-ends have been created so that developers do not directly need to use F-logic or Datalog. Below is example rule using the logic to assign a value to *currency* modifier in context c_3:

$$\forall X : price \; \exists Y : basic \vdash$$
$$X[currency(c_3) \rightarrow Y] \wedge Y[value(c_3) \rightarrow 'KRW'].$$

where variables (e.g., X, Y) are objects, the modifier and attributes of which are represented by methods (which are declared in square brackets). The method *value* is similar to the *value* predicate in context logic of [15]; it returns the ground value of the object in the context specified by the parameter (which is c_3 in the example).

2.3 Characteristics of COIN Lightweight Ontology

A COIN ontology, as shown in Fig. 1, includes only high level concepts (plus their relationships, such as the binary relationships of context modifiers). Thus it is simple and relatively easy to create and reach agreement. But the involved parties do not need to agree on the details of each concept. Each party can continue to use its preferred interpretation for each high level concept. In other words, each party can *conceptually* have its own local ontology. Fig. 2 depicts the conceptual local ontologies for vendors 2 and 3. To avoid clutter, we have omitted attribute names in the figure.

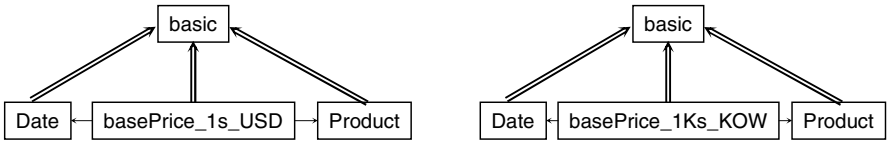


Fig. 2. Conceptual local ontologies for vendor 2 (left) and vendor 3 (right), derivable from COIN lightweight ontology shown in Fig. 1

These local ontologies are not part of the COIN lightweight ontology, but they can be derived from the COIN ontology using the context descriptions. In other words, the COIN lightweight ontology provides a structured way to describe contexts and derive refined local ontologies.

Furthermore, a more traditional global ontology that integrates all the local ontologies could be constructed from the COIN ontology and the accompanying context descriptions. A graphic representation of such a global ontology for the online price comparison example is given in Fig. 3, which includes two intermediate layers (i.e., the layers starting with *BasePrice* and *In USD* concepts, respectively). Concepts

in each layer remove a certain kind of ambiguity. For example, *BasePrice* indicates the kind of price, which does not include shipping and handling charges. The nodes below it further refine the base price concept by specifying the currency, e.g., in USD. Alternatively, the intermediate layers can be omitted. In this case, specialized concepts on the leaf level, such as *basePrice_1s_USD*, directly connect to the generic *Price* concept.

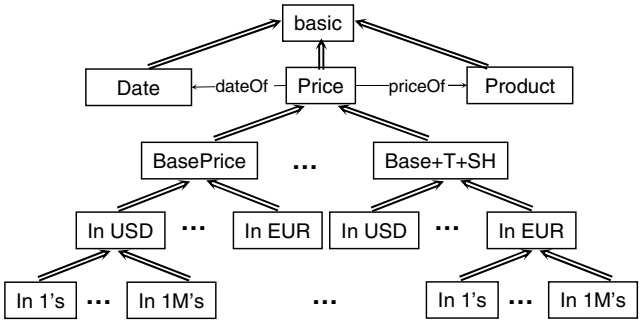


Fig. 3. An example fully-specified global ontology for the online price comparison example. Leaf nodes represents the concepts with specific semantics, e.g., the first leaf node on the left represent the concept of "price, not including taxes or shipping handling, in 1's of USD".

Ontologies are design artifacts. Comparing the artifacts shown in Fig. 1 and Fig. 3, we observe that the COIN approach creates much simpler ontologies – though, for many purposes, they are functionally equivalent. As discussed in [13, 24], the COIN approach has several advantages over the formal ontology approach. First, the COIN ontology is usually much simpler, thus easier to manage. Although in practice it is unlikely that one would create an ontology to include all possible variations (e.g., *basePrice_1M's_USD*), a COIN ontology is still much easier to create than any ontology similar to the one in Fig. 3 even with a smaller number of refined concepts. Second, related to the first point, although the COIN ontology is simple, it provides the means to derive all refined concepts as illustrated in Fig. 3. Third, a COIN ontology facilitates consensus development, because it is relatively easier to agree on a small set of high level concepts than to agree on every piece of detail of a large set of fine-grained concepts. And more importantly, the COIN ontology is much more adaptable to changes. For example, when a new concept "base price + S&H in 1000's of South Korean Won" is needed, the fully specified ontology may need to be updated with insertions of new nodes. The update requires the approval of all parties who agreed on the initial ontology if a single ontology is used, or mappings need to be added to ensure its interoperability with other variants of the *price* concept. In contrast, the COIN approach can accommodate this new concept by adding new context descriptions without changing the ontology. As we will see later, the new mappings may not need to be added when they can be derived from existing mappings using a reasoning mechanism.

The COIN lightweight ontology approach also has advantages over the traditional lightweight ontology approach. Although, similar to the traditional approach, the

COIN ontology does not include detailed descriptions of semantics, it does provide a vocabulary and the structure for describing semantics using context descriptions. As we will see in the next section, the context reasoning mechanism exploits the structure to solve the n^2 problem.

3 Scalable Interoperability with COIN Lightweight Ontology

When data sources and data receivers are in different contexts, conversions (also called lifting rules or mappings) are needed to convert data from source contexts to the receiver context. We call the set of conversions from a context to another context a *composite conversion*. When conversions are specified pair-wise between contexts, it requires ~n^2 composite conversions to achieve interoperability among n contexts. It is costly and error-prone to develop and maintain such a large number of conversions. Thus approaches that hand-code the ~n^2 composite conversions do not scale well when n increases.

The use of lightweight ontology in COIN makes it possible to avoid the above mentioned problem. In addition to using ontology and contexts to represent semantic heterogeneity, COIN also has a reasoning component to determine and reconcile semantic differences. We explain how COIN achieves scalability though conversion composition in the remainder of the section.

3.1 Conversion Composition

In COIN, conversions are not specified as convoluted rules pair-wise between contexts. Instead, they are specified for each modifier between different modifier values. For example, a conversion can be defined for *currency* modifier to convert values in different currencies such as by using an exchange rate function represented by the following predicate:

$$olsen(CurFrom, CurTo, Day, Rate)$$

It returns an exchange *Rate* from *CurFrom* currency to *CurTo* currency on a given *Day*. The function can be implemented externally as a table lookup or as a callable service[1]. We call a conversion defined for a single modifier a *component conversion*.

The component conversions in COIN are also specified using F-logic. Below is an example component conversion for currency modifier; it is parameterized with context C1 and C2 and can convert between any currencies. We use *olsen_* for the skolemized version of original *olsen* predicate.

$$\forall X : price \vdash$$
$$X[cvt(currency, C2) @ C1, u \rightarrow v] \leftarrow$$
$$X[currency(C1) \rightarrow C_f] \wedge X[currency(C2) \rightarrow C_t] \wedge x[dataOf \rightarrow T] \wedge$$
$$olsen\_(A, B, R, D) \wedge C_f \overset{C2}{=} A \wedge C_t \overset{C2}{=} B \wedge T \overset{C2}{=} D \wedge R[value(C2) \rightarrow r] \wedge v = u * r.$$

[1] In many applications using COIN, such conversion functions are implemented by using web wrapped services, such as the www.oanda.com currency conversion web site.

Once all component conversions are defined, composite conversions can be composed automatically using a context reasoning algorithm. Fig. 4 illustrates the concept of conversion composition.

In Fig. 4, the triangle symbol on the left represents the price concept in context c_3, i.e., base price in 1000's of South Korean won (KRW); and the circle symbol on the right represents the price concept in context c_2, i.e., base price in 1's of USD. For data in context c_3 to be viewed in context c_2, they need to be appropriately converted by applying the appropriate composite conversion. The dashed straight arrow represents the application of the composite conversion that would have been implemented manually in other approaches. With the COIN lightweight ontology approach, the composite conversion can be automatically composed using the predefined component conversions. As shown in Fig. 4, we first apply the component conversion for *currency* modifier (represented by $\text{cvt}_{\text{currency}}$), then apply the component conversion for *scaleFactor* modifier (represented by $\text{cvt}_{\text{scaleFactor}}$).

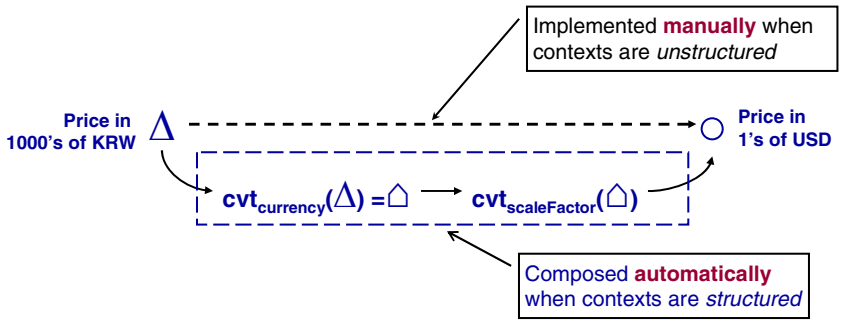


Fig. 4. Composite conversion composed using component conversions. Without composition, one would hand-code a direct conversion to convert the price in 1000's of KRW to the price in 1's of USD; this conversion illustrated by the straight dashed arrow. With COIN, this composite conversion can be derived from the component conversions for currency ($\text{cvt}_{\text{currency}}$) and scale factor ($\text{cvt}_{\text{scaleFactor}}$).

The composition algorithm, shown in Fig. 5, is quite simple. In COIN project, it is implemented in a query rewriting mediator using abductive constraint logic programming (ACLP) [10] and constraint handling rules (CHR) [4]. With the mediator, queries can be issued as if all data sources were in the requester's context (i.e., the target context). The mediator generates mediated queries that contain the composite conversions. Data is converted from source contexts to the requester's context when the mediated queries are executed.

A demonstration of the query mediator is shown in Fig. 6. The source used also includes a *Vendor* column, as shown in the sample schema near the middle of the figure. The source context corresponds to context c_3, and the requester context (c_c_usa2 in the figure) is equivalent to context c_2 in the online price comparison example discussed earlier. In the demonstration, the *QuoteDate* field can have different date formats, which we did not include in the ontology discussed earlier but can be accommodated by adding a *dateFormat* modifier to *Date* concept in the ontology in Fig. 1.

Input: data value V, corresponding concept C in ontology,
source context C1, target context C2
Output: data value V (interpretable in context C2)

```
Find all modifiers of C
  For each modifier mi
       Find and compare mi's values in C1 and C2
       If different: V=cvt_mi(V); else, V=V
Return V
```

Fig. 5. Algorithm for composing composite conversion using component conversions

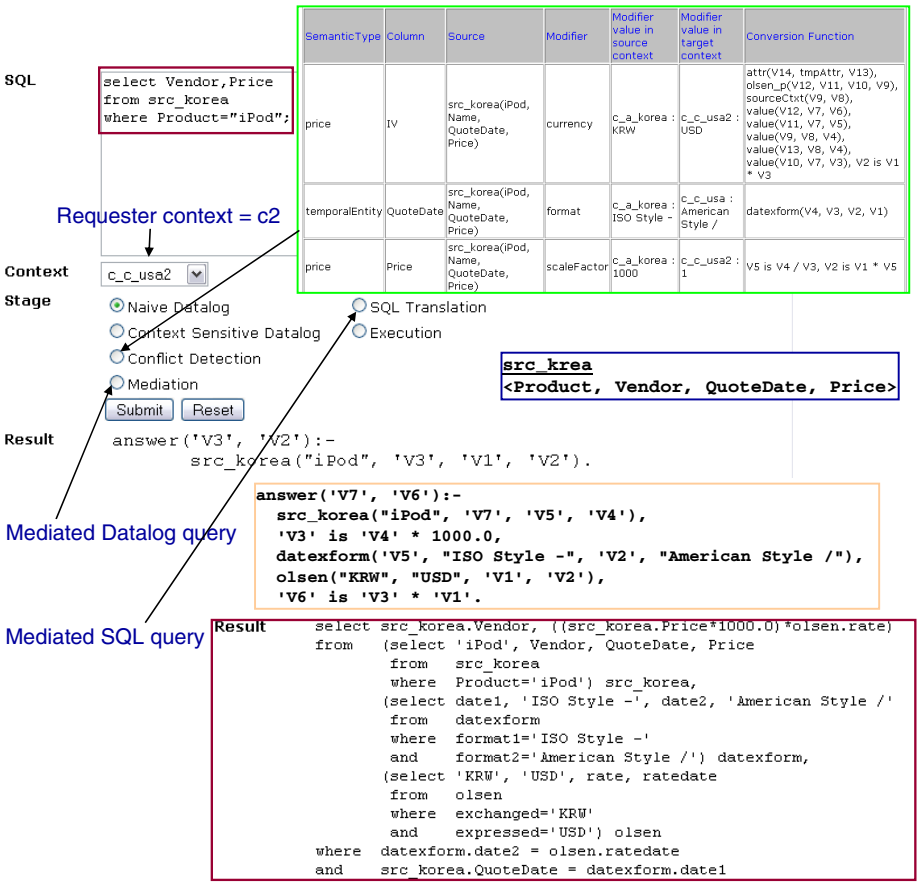


Fig. 6. A demonstration of conversion composition as query mediation

The requester SQL query, shown in the upper left of the figure, need not be aware of any context differences. Our demonstration system allows us to step through the various steps of mediation individually (e.g., converting the SQL to naïve Datalog query, etc.). The Conflict Detection step outputs a table that summarizes the concepts (called Semantic Types) whose modifiers have different values in the source and

requester contexts. A mediated Datalog query is generated using the algorithm shown in Fig. 5. As can be seen, the mediated query contains the necessary conversions to reconcile the context differences (namely currency and scale factor differences of *price* concept, which corresponds to the Price filed in the source table, and format difference of the *Date* concept, which corresponds to the *QuoteDate* field). The mediated Datalog query can be converted an SQL query, which is shown at the bottom in the figure.

3.2 Scalability Benefit

The primary benefit of the composition capability is the small number of component conversions required, thus increased scalability when many data sources and contexts are involved in data integration applications [23, 24].

In the worst case, the number of component conversions required by the light-weight ontology approach of COIN is:

$$\sum_{i=1}^{m} n_i (n_i - 1)$$

where n_i is the number of unique values that the i^{th} modifier has to represent all contexts, m is the number of modifiers in the light-weight ontology.

While the formula appears to be n^2, it is fundamentally different from the approach that supplies *comprehensive conversions* between each pair of contexts. The supplied conversions in COIN are *component conversions,* which are much simpler than the comprehensive conversions that consider the differences of all data elements in all aspects between two contexts. Furthermore, as shown below, the number of component conversions required can be significantly smaller.

Let us use the online price comparison example to illustrate the scalability benefit of the approach. With the given scenario, we can model the 30 unique contexts using the three modifiers in the light-weight ontology shown in Fig. 1. Suppose the number of unique values of each modifier is as shown in Table 2.

Table 2. Modifier values

Modifier	Unique values
currency	10, corresponding to 10 different currencies
scaleFactor	3, i.e., 1, 1000, 1 million
kind	3, i.e., base, base+tax, base+tax+S&H

In the worst case, the light-weight ontology approach needs 102 (i.e., 90+6+6) component conversions. But since the conversions for *currency* and *scaleFactor* modifiers are parameterizable, the actual number of component conversions needed is further reduced to 8, which is a significant improvement from the 870 composite conversions required when conversions are specified pair-wise between contexts.

The number of component conversions can be further reduced when equational relationships exist between contexts with different values of a modifier. Symbolic equation solver techniques have been developed to exploit such relationships [3]. For example, consider the three definitions for price: (A) base price, (B) price with tax

included, and (C) price with tax and shipping & handling included. With known equational relationships among the three price definitions, and two component conversions:

(1) from base_price to base_price+tax (i.e., A to B) and
(2) from base_price+tax to base_price + tax + shipping & handling (i.e., B to C)

the symbolic equation solver can compute the other four conversions automatically (A to C and the three inverses). This technique further reduces the number of component conversions needed for a modifier from $n_i(n_i\text{-}1)$ to $(n_i\text{-}1)$.

In many cases, the component conversion for a modifier can be parameterized, i.e., the component conversion can be applied to convert for any given pair of modifier values. In this case, we only need to supply one component conversion for the modifier, regardless of the number of unique values that the modifier may have. The exchange rate function given earlier is such an example; with it, we only need one component conversion for the *currency* modifier.

We use Fig. 7 to illustrate the intuition of the scalability result.

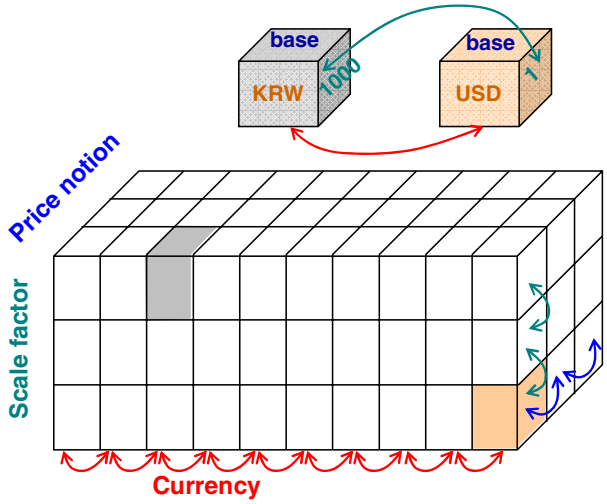


Fig. 7. Intuition of scalability of COIN approach. Component conversions are provided along the modifier axes. Composite conversions between any cubes in the space can be automatically composed.

The modifiers of each ontological concept span a context space within which the variants of the concept exist. Each modifier defines a dimension. In the figure, we show the space spanned by the three modifiers of *price* concept. The component conversions required by the COIN approach are defined along the axes of the modifiers. With the composition capability, the COIN approach can automatically generate all the conversions between units (e.g., the cubes in a three-dimensional space, as sown in Fig. 7) in the space using the component conversions along the dimensions. In contrast, the approaches that suffer from the n^2 problem require the conversions between any two units in the space to be supplied.

4 Related Work and Discussion

The most commonly cited definition for ontology is given in [6], where an ontology is a "formal explicit specification of a share conceptualization". But as discussed in [7, 20], there is not a consensus definition for ontology, and there are many types of ontologies, some of which use formal logic to explicitly capture the intended meanings, and others use a set of mutually agreed terms to provide a shared taxonomy. In the latter case, the intended meanings are not explicitly captured in the ontology, rather, they are implicitly captured in the agreement.

The term *lightweight ontology* has been used very loosely in the literature. Generally speaking, a lightweight ontology refers to a set of concepts organized in a hierarchy with *is_a* relationships. Data dictionaries, product catalogs, and topic maps are often considered to be lightweight ontologies. Opposite to lightweight ontologies are formal ontologies, which often use formal logic to specify constraints, relationships, and other rules that apply to the concepts [8, 14].

The use of ontology and contexts in the COIN approach is quite unique. The ontology provides the necessary structure for context descriptions; and the context descriptions, in turn, disambiguate the high level concepts in the ontology. The structure provided by the ontology also facilitates the provision of component conversions and the automatic composition of composite conversions necessary to enable semantic interoperability among contexts. The resulting solution is scalable because it requires significantly less manually created conversions.

There are other approaches that use ontology or contexts to enable interoperability among disparate data sources [21]. It is beyond the scope of this paper to provide a detailed comparison of these different approaches. We only make comments on a few approaches to further articulate the uniqueness of the COIN approach.

Contexts can be described without using an ontology. For example, they can be described using a context logic [15]. The so described contexts lack the structure like the one provided by the COIN ontology. As a result, a large number of conversions (i.e., lifting rules) are needed to enable semantic interoperability. Below is an example conversion rule to convert price in c_3 to price in c_2 by reconciling the currency and scale factor differences; the rule is a logic implementation of the conversion represented by the straight dashed line in Fig. 4:

$$c_0 : ist(c_2, quote(I, X, D)) \leftarrow ist(c_3, quote(I, P, D)),\quad olsen(krw, usd, D, R), X = P * R * 1000.$$

Suppose there n cubes in the contextual space shown in Fig. 7, the approach requires $n(n\text{-}1)$ conversion rules like the above one to enable full interoperability.

A recent effort tries to categorize lifting rules and attempts to use the patterns revealed to devise general lifting rules [9]. More work is needed to show how these patterns help with creation of general lifting rules and how these rules can be applied to reason with multiple contexts.

Ontology is used in [16], where all types of data level and schema level heterogeneity in multiple data sources are explicitly represented using a semantic conflict resolution ontology (SCROL). For example, when acres and square meters are used in different sources to represent the *area* of a parcel of land, the SCROL ontology will explicitly represent the semantic difference by including two sub-concepts of area: *area_in_acre*, and *area_in_sq_meter*. A SCROL ontology

resembles the one in Fig. 3. The ontology needs to be updated when a new kind of heterogeneity is introduced, e.g., "area in square miles". No characterization on the number of conversions needed is given in the paper.

Ontology is also used in [11] to provide structured context representation for purposes of data interoperability in a multi-database environment. However, we are not certain if their ontology would constitute a lightweight ontology. Nor does the paper provide an assessment about the number of conversions required.

5 Conclusion

The COIN lightweight ontology approach to semantic interoperability has several advantages. The ontology is simple, thus it is easy to create. The semantics of the concepts is described as context descriptions outside the ontology. It can be as a hybrid approach where are a lightweight ontology is annotated with a logic (i.e., F-logic) that can be in a formal ontology approach. The use of modifiers to capture subtle meaning differences provides the structure for describing the subtleties, and facilitates the provision of component conversions, with which any composite conversions can be composed dynamically to reconcile the semantic differences between the sources and the receivers of data.

For future research, we would like to explore the applicability of the COIN approach in other application domains, such as context-aware web services and peer-to-peer information sharing. Another promising area is to apply the context represent-tation and reasoning techniques to Semantic Web applications. Initial work has been done [19] to represent COIN ontology and contexts using Semantic Web languages, such as OWL and RuleML. The preliminary results indicate that COIN lightweight ontology, structured context descriptions, and component lifting rules can be represented using Semantic Web languages. Future work will adapt the reasoning algorithm and evaluate its performance at large scales that are typical on the Semantic Web.

Acknowledgements. This work has been supported, in part, by The MITRE Corporation, the MIT-Malaysia University of Science and Technology (MUST) project, the Singapore-MIT Alliance (SMA), and Suruga Bank.

References

1. Chang, C.H., Kaye, M., Girgis, M.R., Shaalan, K.F.: A Survey of Web Information Extraction System. IEEE Transactions on Knowledge and Data Engineering 18(10), 1411–1428 (2006)
2. Firat, A., Madnick, S.E., Siegel, M.D.: The Cameleon Web Wrapper Engine. In: Workshop on Technologies for E-Services (TES'00), Cairo, Egypt (2000)
3. Firat, A.: Information Integration using Contextual Knowledge and Ontology Merging. In: PhD Thesis, Sloan School of Management. MIT, Cambridge, MA (2003)
4. Frühwirth, T.: Theory and Practice of Constraint Handling Rules. Journal of Logic Programming 37(1-3), 95–138 (1998)
5. Goh, C.H., Bressan, S., Madnick, S., Siegel, M.: Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. ACM Transitions on Information Systems 17(3), 270–293 (1999)

6. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
7. Gruninger, M., Lee, J.: Ontology Applications and Design. Communications of the ACM 45(2), 39–41 (2002)
8. Guarino, N.: Formal Ontology and Information Systems. In: Guarino, N. (ed.) Proceedings of Formal Ontologies in Information Systems (FOIS '98), Trento, Italy, June 6-8, 1998, pp. 3–15. IOS Press, Amsterdam (1998)
9. Guha, R., McCarthy, J.: Varieties of Contexts. In: Blackburn, P., Ghidini, C., Turner, R.M., Giunchiglia, F. (eds.) CONTEXT 2003. LNCS, vol. 2680, pp. 164–177. Springer, Heidelberg (2003)
10. Kakas, A.C., Michael, A., Mourlas, C.: ACLP: Abductive Constraint Logic Programming. Journal of Logic Programming 44(1-3), 129–177 (2000)
11. Kashyap, V., Sheth, A.P.: Semantic and Schematic Similarities between Database Objects: A Context-Based Approach. VLDB Journal 5(4), 276–304 (1996)
12. Kiffer, M., Laussen, G., Wu, J.: Logic Foundations of Object-Oriented and Frame-based Languages. J. ACM 42(4), 741–843 (1995)
13. Madnick, S.E., Zhu, H.: Improving data quality through effective use of data semantics. Data & Knowledge Engineering 59(2), 460–475 (2006)
14. Mädsche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers, Boston, MA (2002)
15. McCarthy, J., Buvac, S.: Formalizing Context (Expanded Notes). In: Aliseda, A., van Glabbeek, R., Westerstahl, D. (eds.) Computing natural language, Sanford University (1997)
16. Ram, S., Park, J.: Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflict. IEEE Transactions on Knowledge and Data Engineering 16(2), 189–202 (2004)
17. Rosenthal, A., Seligman, L., Renner, S.: From Semantic Integration to Semantics Management: Case Studies and a Way Forward. ACM SIGMOD Record 33(4), 44–50 (2004)
18. Seligman, L., Rosenthal, A., Lehner, P., Smith, A.: Data Integration: Where Does the Time Go? IEEE Bulletin of the Technical Committee on Data Engineering 25(3), 3–10 (2002)
19. Tan, P., Madnick, S.E., Tan, K.-L.: Context Mediation in the Semantic Web: Handling OWL Ontology and Data Disparity Through Context Interchange. In: Bussler, C.J., Tannen, V., Fundulaki, I. (eds.) SWDB 2004. LNCS, vol. 3372, pp. 140–154. Springer, Heidelberg (2005)
20. Uschfold, M., Gruninger, M.: Ontologies and Semantics for Seamless Connectivity. ACM SIGMOD Record 33(4), 58–64 (2004)
21. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-Based Integration of Information - A Survey of Existing Approaches. In: IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, pp. 108–117 (2001)
22. Zhu, H., Madnick, S., Siegel, M.: Global Comparison Aggregation Services. In: 1st Workshop on E-Business, Barcelona, Spain (2002)
23. Zhu, H., Madnick, S.E: Context Interchange as a Scalable Solution to Interoperating Amongst Heterogeneous Dynamic Services. In: 3rd Workshop on eBusiness (WEB), Washington, D.C., pp. 150–161 (2004)
24. Zhu, H.: Effective Information Integration and Reutilization: Solutions to Technological Deficiency and Legal Uncertainty. In: Ph.D. Thesis. MIT, Cambridge, MA (2005)

MIT Sloan School of Management

MIT Sloan Working Paper 4558-05
CISL Working Paper 2005-08
October 2005

Improving Data Quality Through Effective Use of Data Semantics (DKE)

Stuart Madnick, Hongwei Zhu

Improving Data Quality Through Effective Use of Data Semantics

Stuart Madnick, Hongwei Zhu

Working Paper CISL# 2005-08

October 2005

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Improving Data Quality Through Effective Use of Data Semantics

Stuart Madnick*, Hongwei Zhu

MIT Sloan School of Management, 30 Wadsworth Street, E53-320, Cambridge, MA 02142, USA

Abstract

Data quality issues have taken on increasing importance in recent years. In our research, we have discovered that many "data quality" problems are actually "data misinterpretation" problems – that is, problems caused by heterogeneous data semantics. In this paper, we first identify semantic heterogeneities that, when not resolved, often cause data quality problems. We discuss the especially challenging problem of aggregational ontological heterogeneity, which concerns how complex entities and their relationships are aggregated. Then we illustrate how COntext INterchange (COIN) technology can be used to capture data semantics and reconcile semantic heterogeneities, thereby improving data quality.

Keywords: Data Quality, Data Semantics, Semantic Heterogeneity, Ontology, Context

1. Introduction

Data quality issues have taken on increasing importance in recent years. In our research, we have discovered that many "data quality" problems are actually "data misinterpretation" problems – that is, problems with data semantics. To illustrate how complex this can become, consider Fig. 1. This data summarizes the P/E ratio for DaimlerChrysler obtained from four different financial information sources – all obtained on the same day within minutes of each other. Note that the four sources gave radically different values for P/E ratio.

Source	P/E Ratio
ABC	11.6
Bloomberg	5.57
DBC	19.19
MarketGuide	7.46

Fig. 1. P/E ratios for DaimlerChrysler.

The obvious questions to ask are: "Which source is correct?" and "Why are the other sources wrong – i.e., of bad data quality?" The possibly surprising answer is: they are all correct!

The issue is, what do you really mean by "P/E ratio"[1]. The answer lies in the multiple interpretations and uses of the term "P/E ratio" in financial circles. The earnings are for the entire year in some sources but in one source are only for the last quarter. Even when earnings are for a full year, are they:

- the last 12 months?

* Corresponding author. Tel: +1 617 253 6671; fax: +1 617 253 3321.
Email addresses: smadnick@mit.edu (S. Madnick), mrzhu@mit.edu (H. Zhu).

[1] Some of these sites even provide a glossary which gives a definition of such terms and they are very concise in saying something like "P/E ratio" is "the current stock price divided by the earnings". As it turns out, this does not really help us to explain the differences.

- the last calendar year?
- the last fiscal year? or
- the last three historical quarters and the estimated current quarter – a popular usage?

Such information, which we call *context*, is often not explicitly captured in a form that can be used by the query answering system to reconcile semantic differences in data from different sources. Serious consequences can result from not being aware of the differences in contexts and data semantics. Consider a financial trader that used DBC to get P/E ratio information yesterday and got 19.19. Today he used Bloomberg and got 5.57 (low P/E's usually indicate good bargains) – thinking that something wonderful had happened he might decide to buy many shares of DaimlerChrysler today. In fact, nothing had actually changed, except for changing the source that he used. It would be natural for this trader (after possibly losing a significant amount of money due to this decision) to feel that he had encountered a data quality problem.

We would argue that what appeared to be a data quality problem is actually a data misinterpretation problem. The data source did not have any "error," the data that it provided was exactly the data that it intended to provide – it just did not have the meaning that the receiver expected. In other words, the issue is not what is right or wrong, it is about how data in one context can be used in a different context.

Before going any further, it should be noted that if all sources and all receivers of data always had the exact same meanings, this problem would not occur. This is a desirable goal – one frequently sought through standardization efforts. But these standardizations are often unsuccessful for many reasons [18], e.g., there are legitimate needs for representing and interpreting data in different ways to suit different purposes[2]. This creates the well known problem of semantic heterogeneities that exist pervasively in information systems. It is crucial that we understand the kinds of heterogeneity and develop technologies to provide data that is consistent with receiver preference, thereby improve the data quality at the receiver end. Such a solution can have significant impact as the estimated cost of information mishandling in businesses worldwide is tremendous [19].

In the next section, we exemplify the semantic heterogeneities that, when not reconciled, can cause data quality problems. Then, we present the Context Interchange technology and show how it can be used to capture data semantics and dynamically reconcile semantic differences between the sources and the receivers. This technology supports the uniformity required by any specific receiver, at same time, it supports heterogeneity by preserving the autonomy of all sources and receivers. We conclude in the last section and point out directions for future research.

2. Heterogeneous Data Semantics

There have been a number of studies that identify and catalog various semantic heterogeneities [3,11,16,17]. A subset of the heterogeneities are related to data quality that we address in this paper and can be categorized into two main groups: (1) representational heterogeneity and (2) ontological heterogeneity. Data semantics can sometimes change over time; therefore, representational and ontological semantics of a source or a receiver can evolve, resulting in temporal semantic heterogeneities. These categories are summarized in Fig. 2 and explained next.

[2] A full discussion of all the difficulties with standardization is beyond the scope of this paper. It is worth noting that the "Treaty of the Meter" committing the U.S. government to go metric was initially signed in 1875.

Representational	Ontological
Temporal	

	Snapshot example	Temporal example
Representational	*Currency*: EUR in source *v.* USD in receiver	*Currency*: DEM until 12/31/98 EUR since 1/1/99
Ontological	*Profit*: Net excl. tax in source *v.* Gross incl. tax in receiver	*Profit*: Net until 1999 Gross since 2000

Fig. 2. Data quality related semantic heterogeneities.

Representational heterogeneity – The same concept can have different representations in different sources and receivers. For example, the day of *March 4, 2005* can be represented as *03/04/05*, *05-03-04*, etc; packaging dimensions can be expressed in metric units or in English units; price data can be quoted in different currencies and using different scale factors.

Temporal Representational heterogeneity – The representation in a source or a receiver can also change. For example, a price database in Turkey may list prices in millions of Turkish liras (TRL), but after the Turkish New Lira (TRY) was introduced on January 1, 2005, it may start to list prices in unit of Turkish New Lira[3].

Ontological heterogeneity – The same term is often used to refer to similar but slightly different concepts. Known and quantifiable relationships often exist amongst these concepts. We have already seen an example of this regarding the multiple interpretations of "P/E ratio" in Fig. 1.

Temporal Ontological heterogeneity – In addition, in the same source or receiver, the meaning of a term can shift over time, often due to changes of needs or requirements. For example, *profit* can refer to *gross profit* that includes all taxes collected on behalf of government, or *net profit* that excludes those taxes. Net profit can be calculated from gross profit by deducting the taxes, and vice versa. The "Profit" field in a database may refer to net profit at one time and refer to gross profit at another, because of changes in reporting rules.

Aggregational Ontological heterogeneity – Another variation can be that the profit of a firm may include that from majority owned subsidiaries in one case, and excludes them in another case (possibly due to different reporting rules in different countries or for different purposes.) Aggregational ontological heterogeneity has to do with what is included/aggregated in the meaning of an entity or a relationship. A specific example of this situation, sometimes called corporate housekeeping, will be presented later.

Representational and ontological data semantics is often embedded in the explicit data and the implicit assumptions; semantic heterogeneities exist when the implicit assumptions in the sources do not match the implied expectations of the receivers. They must be reconciled to ensure the correct interpretation of the data by the receivers. In the following, we will use several examples to illustrate the semantic heterogeneities and their effects on data quality.

Example 1: Temporal Representational Semantics (Yahoo Historical Stock Prices)

When the same company stock is traded at different stock exchanges around the world, there may be *small* price differences between exchanges, creating arbitraging opportunities (i.e., buying low in one place and selling high at another). Fig. 3 gives an example of how big the

[3] The following fact may help explain why this could be case: 1 USD = 1.39 million TRL; 1 TRY = 1 million TRL; it would be cumbersome to list many 0's if prices were listed in unit of TRL.

differences can be – on the left are IBM stock prices in Frankfurt, Germany, on the right are that in New York, USA. We notice that the values between the two exchanges during the same time period are *huge* (comparing the values in the brackets); in addition, there is an abrupt price drop in Frankfurt while the prices in New York are stable (comparing the values in the circles). This is quite unusual! Again, one may start to question about data quality in the sources, but in fact, the peculiarities in the data are due to semantic mismatches – the implied currencies not only differ between the two exchanges, but also changed in Frankfurt from Deutschmark (DEM) to Euro (EUR); the currency in New York has always been USD. This is an example of representational heterogeneity that also evolves over time. ■

Frankfurt, Germany

PRICES

Date	Open	High	Low	Close	Volume	Adj Close*
8-Jan-99	162.20	165.00	162.20	163.00	5,220	78.96
7-Jan-99	162.00	162.50	160.00	160.00	3,647	77.51
6-Jan-99	160.90	164.00	160.90	164.00	23,616	79.45
5-Jan-99	154.00	155.00	154.00	155.00	5,975	75.09
4-Jan-99	155.00	158.00	155.00	155.50	7,024	75.33
30-Dec-98	311.50	311.50	309.00	309.00	28,324	149.69
29-Dec-98	314.90	314.90	313.30	314.00	22,313	152.12
28-Dec-98	314.80	314.80	314.00	314.00	26,640	152.12
23-Dec-98	303.50	305.00	302.50	305.00	38,363	147.76
22-Dec-98	293.60	297.50	293.00	294.00	29,899	142.43
21-Dec-98	282.50	293.00	282.50	293.00	27,603	141.94

* Close price adjusted for dividends and splits.

New York, USA

PRICES

Date	Open	High	Low	Close	Volume	Adj Close*
8-Jan-99	191.00	192.00	185.63	187.56	4,593,100	89.99
7-Jan-99	187.94	192.38	187.00	190.19	4,155,300	91.25
6-Jan-99	190.31	192.75	188.50	188.75	4,775,300	90.56
5-Jan-99	183.00	189.88	182.81	189.63	4,955,900	90.98
4-Jan-99	185.00	186.50	181.50	183.00	4,077,500	87.80
31-Dec-98	186.75	187.19	183.50	184.38	1,933,800	88.46
30-Dec-98	186.88	188.63	186.31	186.75	2,410,300	89.60
29-Dec-98	188.63	188.94	187.00	187.13	1,881,700	89.78
28-Dec-98	186.50	189.94	186.00	189.25	2,637,200	90.80
24-Dec-98	184.75	187.94	184.06	187.94	1,527,700	90.17
23-Dec-98	182.69	185.38	181.13	185.00	3,537,800	88.76
22-Dec-98	177.50	183.00	175.25	182.25	4,353,600	87.44
21-Dec-98	171.56	178.94	171.56	176.38	3,744,500	84.62

* Close price adjusted for dividends and splits.

Fig. 3. IBM stock prices at different exchanges (from Yahoo).

Example 2: Aggregational Ontological Semantics (Corporate Householding)

The rapidly changing business environment has witnessed widespread and rapid changes in *corporate structure* and *corporate relationships*. Regulations, deregulations, acquisitions, consolidations, mergers, spin-offs, strategic alliances, partnerships, joint ventures, new regional headquarters, new branches, bankruptcies, franchises … all these make understanding corporate relationships an intimidating job. Moreover, the same two corporation entities may relate to each other very differently when marketing is concerned than when auditing is concerned. That is, interpreting corporate structure and corporate relationships depends on the task at hand. To understand the challenges, let us consider some typical, simple, but important questions that an organization, such as IBM or MIT, might have about their relationships:

[MIT]: "How much did we buy from IBM this year?"

[IBM]: "How much did we sell to MIT this year?"

The first question frequently arises in the Procurement and Purchasing departments of many companies, as well as at more strategic levels. The second question frequently arises in the Marketing departments of many companies and is often related to Customer Relationship Management (CRM) efforts, also at more strategic levels. Logically, one might expect that the answers to these two questions would be the same – but frequently they are not, furthermore one often gets multiple different answers even within each company.

These types of questions are not limited to manufacturers of physical goods, a financial services company, such as Merrill Lynch, might ask:

[Merrill Lynch]: "How much have we loaned to IBM?"

[IBM]: "How much do we owe Merrill Lynch?"

On the surface, these questions may sound like both important and simple questions to be able to answer. In reality, there are many reasons why they are difficult and have multiple differing answers.

At least three types of challenges must be overcome to answer questions such as the ones illustrated above: (a) representational semantic heterogeneity, (b) entity aggregational ontological heterogeneity, and (c) relationship aggregational ontological heterogeneity. The first two concern *what IBM or MIT is*, and the third one concerns *how IBM and MIT are related*. These challenges provide a typology for understanding what is sometimes called the Corporate Householding, as illustrated in Fig. 4 and explained below.

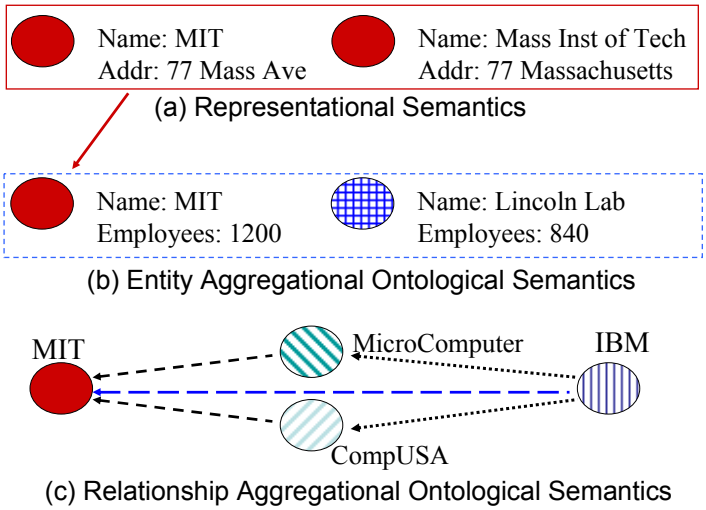


Fig 4. Typology for Corporate Householding Challenges.

(a) *Representational Semantics*. In general, there are rarely complete unambiguous universal identifiers for either people or companies. Two names may refer to the same physical entity even though they were not intended to create confusions in the beginning. For example, the names "James Jones", "J. Jones", and "Jim Jones" might appear in different databases, but actually be referring to the same person. The same problems exist for companies. As shown in Fig. 4(a), the names "MIT", "Mass Inst of Tech", "Massachusetts Institute of Technology", and many other variations might all be used to refer to the exact same entity. They are different simply because the users of these names choose to do so. Thus, we need to be able to identify the same entity correctly and efficiently when naming confusion happens. This problem has also been called Identical Entity Instance Identification [10]. That is, the same identical entity might appear as multiple instances (i.e., different forms) – but it is still the same entity.

(b) *Entity Aggregational Ontological Semantics*. Even after we have determined that "MIT", "Mass Inst of Tech", "Massachusetts Institute of Technology" all refer to the same entity, we need to determine what exactly is that entity? That is, what other unique entities are to be included or aggregated into the intended definition of "MIT." For example, the MIT Lincoln Lab, according to its home page, is "the Federally Funded Research and Development Center of the Massachusetts Institute of Technology." It is located in Lexington and physically separated from the main campus of MIT (sometimes referred to as the "on-campus MIT"), which is in

Cambridge. Lincoln Lab has a budget of about $500 million, which is about equal to the rest of MIT.

Problem arises when people ask questions such as "How many employees does MIT have?" or "How much was MIT's budget last year?". In the case illustrated in Fig. 4(b), should the Lincoln Lab employees or budget be included in the "MIT" calculation and in which cases they should not be? Under some circumstances, the MIT Lincoln Lab number should be included, whereas under other circumstances they should not be. We refer to these differing circumstances as different contexts. To know which case applies under each category of circumstances, we must know the context. As noted earlier, we refer to this type of problem as Entity Aggregational Ontological heterogeneity.

(c) *Relationship Aggregational Ontological Semantics*. Furthermore, even after we have resolved the aggregation of entities, we still need to determine the relationships between the entities. As illustrated in Fig. 4(c), the buying/selling relationships between MIT and IBM can be direct or indirect through other channels. Consider our original question – for IBM: "How much did we sell to MIT this year?" The answer to question varies depending on the aggregation of sales channels. For example, under some circumstances, only the direct sales from IBM to MIT are included, whereas under other circumstances, sales through other channels (e.g., through partners, retailers, etc.) are also included.

In summary, the answers to the questions can be dramatically different because of the multiple situations that exist. Different answers do not signify that some answers are wrong; all answers can be correct under their corresponding circumstances, i.e., in their own contexts.■

Example 3: Temporal Ontological Semantics (Code v. What Code Denotes)

In everyday communications and in various information systems, it is very common that we refer to things using various codes, e.g., product codes of a company, subject numbers in a university catalog, and ticker symbols commonly used to refer to company stocks. Codes are sometimes reused in certain systems, thus the same code can denote different things at different times. For example, subject number "6.891" at MIT has been used to denote "Multiprocessor Synchronization", "Techniques in Artificial Intelligence", "Computational Evolutionary Biology", and many other subjects in the past decade. As an another example, ticker symbol "C" used to be the symbol for Chrysler; after Chrysler merged with Daimler-Benz in 1997, the merged company chose to use "DCX"; on December 4, 1998, the symbol "C" was assigned to Citigroup, which was listed as "CCI" before this change. ■

Example 4: Temporal Aggregational Ontological Semantics (Yugoslavia)

To study the economic and environmental development of different parts of the world, one often needs longitudinal data from various sources. In the past 30 years, certain regions have gone through significant restructuring, e.g., one country breaking up into several countries. Such dramatic changes can make it difficult to use data from multiple sources or even from a single source. As an example, suppose a Balkans specialist is interested in studying the CO_2 emissions in the region of former *Yugoslavia* during 1980-2000 and prefers to refer to the region (i.e. the geographic area of the territory of former Yugoslavia) as Yugoslavia. Data sources like the Carbon Dioxide Information Analysis Center (CDIAC)[4] at Oak Ridge National Laboratory organize data by country. Fig. 5 lists some sample data from CDIAC. Yugoslavia as a country, whose official name is *Socialist Federal Republic of Yugoslavia* in 1963-1991, was broken into

[4] http://cdiac.esd.ornl.gov/home.html

five independent countries in 1991: *Slovenia*, *Croatia*, *Macedonia*, *Bosnia and Herzegovina*, and *Federal Republic of Yugoslavia* (also called *Yugoslavia* for short in certain other sources). Suppose prior to the break-up the specialist had been using the following SQL query to obtain data from the CDIAC source:

```
Select CO2Emissions from CDIAC where Country = "Yugoslavia";
```

Before the break-up, "Yugoslavia" in the receiver coincidentally referred to the same geographic area as to what "Yugoslavia" in the source referred, therefore, the query worked correctly for the receiver until 1991. After the break-up, the query stopped working because no country is named "Yugoslavia" (or had the source continued to use "Yugoslavia" for the Federal Republic of Yugoslavia, the query would return wrong data because "Yugoslavia" in the source and the receiver refer to two different geographic areas). ■

Country	**Year**	**CO2Emissions**
...	...	...
Yugoslavia	1990	35604
Yugoslavia	1991	24055
Slovenia	1992	3361
Croatia	1992	4587
Macedonia	1992	2902
Bosnia-Herzegovinia[5]	1992	1289
Federal Republic of Yugoslavia	1992	12202
...	...	...

Fig. 5. Sample CO2 emissions data from CDIAC.

These examples demonstrate that poor data quality can result from representational and ontological heterogeneities between the sources and the receivers. They also suggest that we can improve data quality by resolving these heterogeneities. In simple cases, this can be done manually by the receivers. But in most practical cases that involve a large number of sources and data elements, a manual reconciliation will be difficult and error prone. In the next section, we will introduce the Context Interchange technology and show how it is used to improve data quality by automatically reconciling semantic differences between the sources and the receivers.

3. Improving Data Quality with Context Interchange Technology

3.1. Context Interchange Overview

COntext INterchange (COIN) [7,9,10] is a knowledge-based mediation technology that enables meaningful use of heterogeneous databases where there are semantic differences. With the COIN technology, a user (i.e., information receiver) is relieved from keeping track of various source contexts and can use the sources as if they were in the user context. Semantic differences are identified and reconciled by the mediation service of COIN. The overall COIN system includes not only the mediation infrastructure and services, but also a wrapping technology and middleware services for accessing the source information and facilitating the integration of the mediated results into end-user applications (see Fig. 6). The wrappers are physical and logical gateways providing a uniform access to the disparate sources over the network [5].

[5] Correct spelling is "Herzegovina", which is an error; we do not address this kind of data quality problem in this paper.

The set of Context Mediation Services comprises a Context Mediator, a Query Optimizer, and a Query Executioner. The Context Mediator is in charge of the identification and resolution of potential semantic differences induced by a query. This automatic detection and reconciliation of differences present in different information sources is made possible by accessing the knowledge of the underlying application domain, as well as informational content and implicit assumptions associated with the receivers and sources. These bodies of declarative knowledge are represented in the form of a shared ontology, a set of elevation axioms, and a set of context definitions, which we explain below.

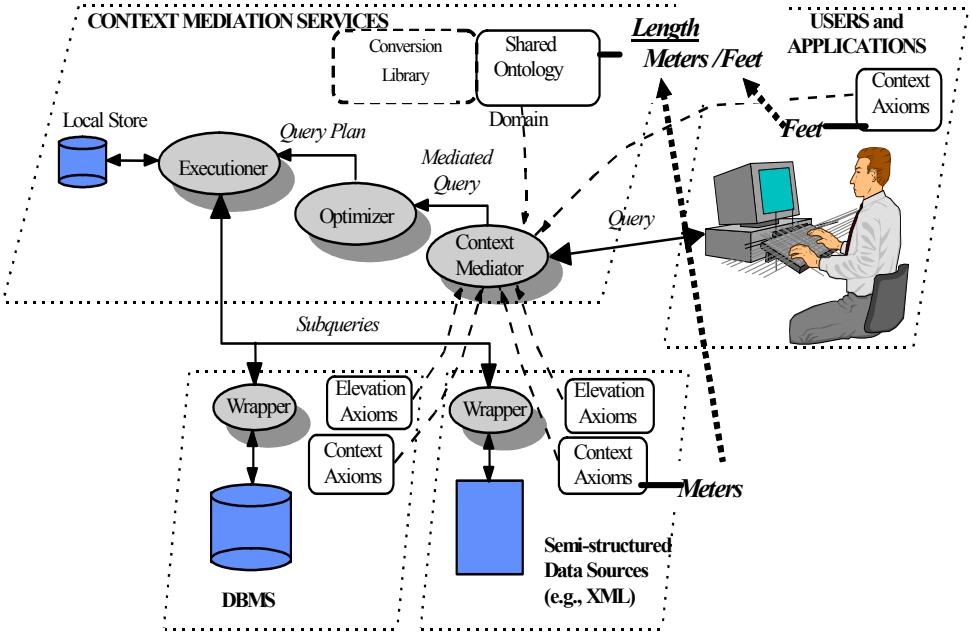


Fig. 6. The architecture of the context interchange system.

The input to the mediator is a user query assuming that all sources were in the user context. The result of the mediation is a mediated query that includes the instructions on how to reconcile the semantic differences in different contexts involved in the user query. To retrieve the data from the disparate information sources, the mediated query is then transformed into a query execution plan, which is optimized, taking into account the topology of the network of sources and their capabilities. The plan is then executed to retrieve the data from the various sources.

For the mediator to identity and reconcile semantic difference, necessary knowledge about data semantics needs to be formally represented. For purposes of knowledge representation, COIN adopts an object-oriented logic data model, based on the formal theory of F-Logic [13], a first order logic with syntactic sugar to support object-orientation (e.g., inheritance, polymorphism, etc.). Loosely speaking, the COIN data model has three elements, for which we give a brief overview below and provide further explanations in the next sub-section:

- The Shared Ontology is a collection of concepts, also called rich types or semantic types, that define the domain of discourse (e.g., "Length");
- Elevation Axioms for each source identify the semantic objects (instances of semantic types) corresponding to source data elements, define integrity constraints, and specify general properties of the sources;

- Context Descriptions annotate the different interpretations of the semantic objects in the different sources or from a receiver's point of view (e.g., "Length" might be expressed in "Feet" or "Meters").

Finally, there is a conversion library which provides conversion functions for resolving potential semantic differences. The conversion functions can be defined declaratively or can use external services or external procedures. The relevant conversion functions are gathered and composed during mediation to resolve the differences. No global or exhaustive pair-wise definition of the conflict resolution procedures is needed. The mediator is implemented using abductive constraint logic programming (ACLP) [12], which not only rewrites queries to reconcile semantic differences, but also performs semantic query optimization.

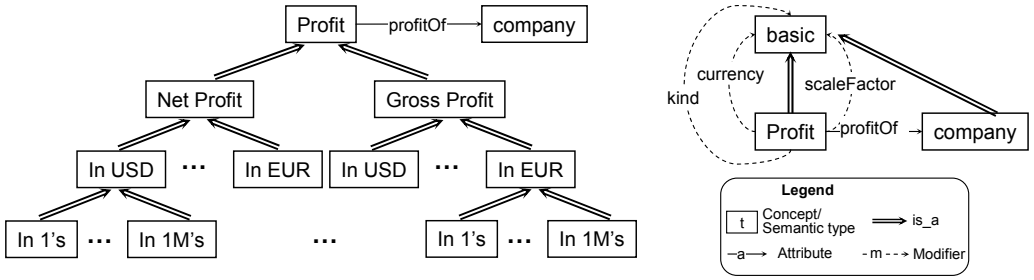
3.2. Representing Heterogeneous Semantics using Ontology and Contexts

To a certain extent, ontology modeling and entity-relationship modeling [2,4] share the same objective of providing a formal way of representing things in the real world. An ontology usually consists of a set of terms corresponding to a set of predefined concepts (similar to entities), relationships between concepts, and certain constraints. There are two types of binary relationships between concepts: *is_a*, and *attribute*. The *is_a* relationship indicates that a concept is more specific (or conversely, more general) than another (e.g., the concept of *net profit* is more specific than the concept of *profit*); the *attribute* relationship simply indicates that a concept is an attribute of another (e.g., the *company* concept is the *profit_of* attribute of the *profit* concept).

A high level concept can have various specializations. As shown in Fig. 7(a) below, *profit* can have specializations such as *gross profit* and *net profit*, each can be further specialized to use various currencies, which can be further specialized to use different scale factors (e.g., in thousands or millions). Since the purpose of ontology is to share knowledge, it is tempting to fully describe these specializations in the ontology so that there will be no ambiguity in the semantics of the concepts. However, the ontology of this approach is difficult to develop because (1) the ontology often consists of a large number of concepts, and (2) it requires various parties engaged in knowledge sharing to agree on the precise definitions of each concept in the ontology.

The COIN ontology departs from the above approach, as shown in Fig.7(b). It only requires the parties to agree on a small set of general concepts. Detailed definitions (i.e., specializations) of the general concepts are captured outside the ontology as localized context descriptions. The context descriptions usually correspond to the implicit (and sometimes evolving) assumptions made by the data sources and receivers. To facilitate context description, the COIN ontology includes a special kind of attribute, called the *modifier*. Contexts are described by assigning values to modifiers.

These two different approaches are illustrated in Fig. 7 using the company profit example.



(a) a fully specified ontology of *profit* (b) a COIN ontology of *profit*

Fig 7. Fully specified ontology v. COIN ontology.

The fully specified ontology in Fig. 7(a) contains all possible variations/specializations of the concept *profit*, organized in a multi-level and multi-branch hierarchy. Each leaf node represents a most specific *profit* concept. For example, the leftmost node at the bottom represents a profit concept that is a "net profit in 1's of USD". In contrast, the COIN ontology contains only concepts in higher levels (e.g, *profit*), further refinements of these concepts do not appear in the ontology; rather, they are specified outside the ontology and are described using modifiers (e.g., if in a context, the profit data is "net profit in 1's USD", the context can be described by assigning appropriate values to the modifiers, i.e., *kind*="net", *scaleFactor*= "1", and *currency*="USD").

Compared with the fully specified approach, the COIN approach has several advantages. First, a COIN ontology is usually much simpler, thus easier to manage. Second, it facilitates consensus development, because it is relatively easier to agree on a small set of high level concepts than to agree on every piece of detail of a large set of fine-grained concepts. And more importantly, a COIN ontology is much more adaptable to changes. For example, when a new concept "*net profit* in billions of South Korean Won" is needed, the fully specified ontology needs to be updated with insertions of new nodes. The update requires the approval of all parties who agreed on the initial ontology. In contrast, the COIN approach can accommodate this new concept by adding new context descriptions without changing the ontology.

Another important distinction is in the provision of conversions for reconciling semantic differences. Other approaches tend to provide pair-wise conversions between the data elements that correspond to the leaf nodes in the fully-specified ontology, e.g., a conversion between the data of "net profit in 1's of USD" and the data of "gross profit in millions of EUR". We call such conversions *composite conversions*. In the COIN approach, conversions are provided for each modifier; such conversions are called *component conversions*. All pair-wise composite conversions are automatically composed by the mediator using the component conversions. In the example illustrated in Fig.7, with three component conversions (i.e., one for each modifier), the COIN mediator can compose all composite conversions as needed between any pair of the leaf nodes in the fully specified ontology.

With the ontological constructs and the component conversions in the COIN approach, all data quality related semantic heterogeneities identified in the previous section can be represented and processed. We will use the simplified ontology for the Yahoo historic stock price example, shown below in Fig. 8, to explain how this is done.

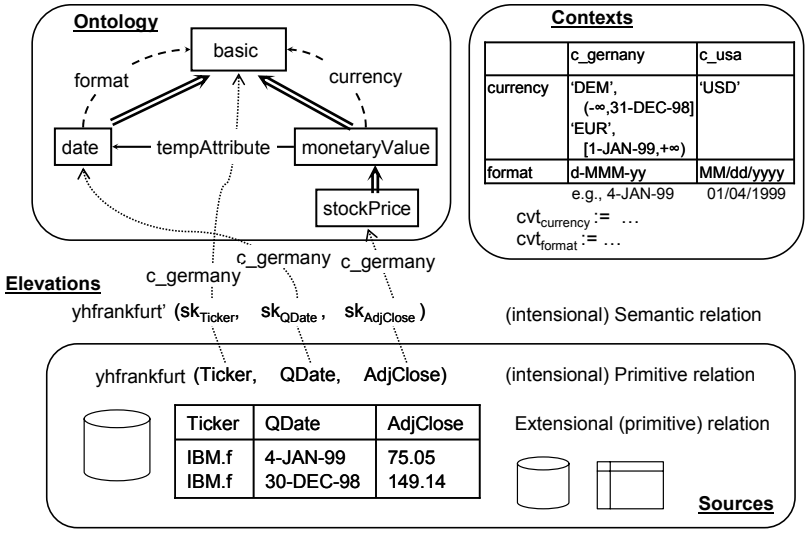


Fig. 8. COIN ontology, contexts, and elevations for historical stock price example.

Ontology and contexts are shown in the upper half of Fig. 8. There are two modifiers in the ontology: *format* for describing different date formats, and *currency* for describing different currencies. We show two sample contexts: (1) *c_germany* for the Yahoo site that provides stock prices at Frankfurt Stock Exchange (we call the source *yhfrankfurt*); and (2) *c_usa* for receivers, say in the U.S. We describe contexts by assigning values to the modifiers in the ontology, as shown in the upper-right corner of the figure.

Formally, we use F-Logic[6] formulas (sometimes called rules) to describe contexts. Temporal semantics can be described using multi-valued modifiers, i.e., a modifier can have different values in different time periods within a context, and we call such context a *temporal context*. For example, to describe that in context *c_germany*, the currency is "DEM" until December 31, 1998, and is "EUR" since January 1, 1999, we use the following two rules to assign different values to modifier *currency* in the two respective time periods:

$$
\begin{aligned}
&O : monetaryValue, currency(c\_germany, O) : basic \vdash \\
&\quad O[currency@c\_germany \rightarrow currency(c\_germany, O)] \leftarrow \\
&\qquad currency(c\_germany, O)[value@c\_germany \rightarrow "DEM"], \\
&\qquad rcvContext(C), O[tempAttribute \rightarrow T], T[value@C \rightarrow T_v], \\
&\qquad skolem(date, "31\text{–}DEC\text{–}98", c\_germany, 1, cste("31\text{–}DEC\text{–}98"))[value@C \rightarrow T_c] \\
&\qquad T_v \le T_c .
\end{aligned}
$$

$$
\begin{aligned}
&O : monetaryValue, currency(c\_germany, O) : basic \vdash \\
&\quad O[currency@c\_germany \rightarrow currency(c\_germany, O)] \leftarrow \\
&\qquad currency(c\_germany, O)[value@c\_germany \rightarrow "EUR"], \\
&\qquad rcvContext(C), O[tempAttribute \rightarrow T], T[value@C \rightarrow T_v], \\
&\qquad skolem(date, "1\text{–}JAN\text{–}99", c\_germany, 1, cste("1\text{–}JAN\text{–}99"))[value@C \rightarrow T_c] \\
&\qquad T_v \ge T_c .
\end{aligned}
$$

Readers are referred to [13] for the details of the F-logic language; here we provide a brief explanation on how it is used for context descriptions. A modifier is represented as a parameterized *method* of an object and expressed within the square brackets following the object.

[6] Although F-Logic is the internal representation used with COIN, a user-friendly interface make it unnecessary for any user (either context administrator or query user) to know F-Logic.

For example, in the head of the first rule above, *currency* modifier is the currency method of the semantic object *O*, whose type is *monetaryValue*. Given the parameter *c_germany*, the method returns an object, of type *basic*, represented by a *Skolem* function (also called a *Skolem* object) with *c_germany* and *O* as the parameters. The body of the rule (after "←") indicates that the *value* of the *Skolem* object in context *c_germany* is "DEM" when the *tempAttribute* attribute of *O* is before "31-DED-98".

The *currency* modifier in context *c_usa* and the *format* modifier in both contexts can be specified similarly. These specifications are simpler because the modifier value is not time dependent. For example, the rule below states that in context *c_usa* the currency is "USD":

$$O : monetaryValue, currency(c\_usa, O) : basic \vdash$$
$$O[currency@c\_usa \rightarrow currency(c\_usa, O)] \leftarrow$$
$$currency(c\_usa, O)[value@c\_usa \rightarrow "USD"].$$

Component conversions for modifiers *format* and *currency* are specified using F-logic rules, as well:

$$D : date \vdash$$
$$D[cvt@format, C_r, M_s, M_r, U \rightarrow V] \leftarrow$$
$$datecvt(U, M_s, V, M_r).$$

$$M : monetaryValue \vdash$$
$$M[cvt@currency, C_r, M_s, M_r, U \rightarrow V] \leftarrow$$
$$M[tempAttribute \rightarrow T], T[value@C_r \rightarrow T_v], olsen'(F_c', T_c', R', D'),$$
$$F_c'[value@C_r \rightarrow M_s], T_c'[value@C_r \rightarrow M_r], D'[value@C_r \rightarrow T_v],$$
$$R'[value@C_r \rightarrow R_v], V = U * R_v.$$

Both rules use external programs/services. The first rule uses the external program *datecvt* to perform date format conversions; the second rule uses the external service *olsen* to obtain the exchange rate between a pair of currencies on a given day. Wrappers [5] are used for these external programs/services so that they can be accessed like relational databases.

In the lower half of Fig. 8, we show the data source *yhfrankfurt* with its schema and two sample records. The elevation axioms map each column of a relation to a concept in the ontology and associate each column with a context. Thus, for each relation (which we call *primitive relation*) in the source there is a *semantic relation*. Each attribute in a semantic relation is a (meta-) semantic object (i.e., an instance of a semantic type), which has access to the context descriptions and component conversions defined for the modifiers of the corresponding semantic type.

3.3. Reconciling Semantic Heterogeneities Through Mediation

Once contexts are recorded for all sources and receivers, and component conversions are provided, a receiver can query any collection of sources as if all they were in the receiver context. The mediator will intercept the query, compare the contexts involved, introduce appropriate component conversions, and generate a mediated query that reconciles the semantic differences.

The implementation of the mediator is based on the formal theory of abductive constraint logic programming (ACLP) [12], where abductive inference is interleaved with concurrent constraint solving. The constraint store collects all abducible predicates generated by abductive inference. All abducible predicates are treated as constraints, the consistency of which is handled by constraint solvers defined using the declarative language Constraint Handling Rules (CHR) [8]. For example, the descriptions of a temporal context often involve comparisons of time

values; when these comparison predicates are abduced, they are treated as constraints and processed by temporal constraint solvers, which generates a common time period during which all involved modifier are singly valued. We also use CHR to solve symbolic equations [6] and perform semantic query optimization. Detailed descriptions of the implementation can be found in [7,9,10].

Below, we use the Yahoo historical stock price example to illustrate how COIN is used to provide meaningful data to the receiver without the receiver being burdened to keep track of semantic heterogeneities.

Suppose a receiver in context *c_usa* wants to retrieve historical stock prices at Frankfurt Stock Exchange. The receiver prefers to see the adjusted close price in USD and the date in MM/dd/yyyy format (e.g., 01/10/1999). Using the web wrapper technology, we can superimpose the following relational schema to the data source at Yahoo Finance website:

```
YHFrankfurt<Ticker,QDate,AdjClose,
            StartMonth,StartDay,StartYear,EndMonth,EndDay,EndYear>
```

where the last six attributes corresponds to the month, day, and year of "Start Date" and "End Date". These attributes are necessary only because the source is not able to accept date range specified as inequalities on the "QDate" attribute; instead, it is only able to accept equalities on the last six attributes. Like semantic differences, such capability differences should be processed by the system, not the receivers. Therefore, we let the source expose a much simpler schema:

```
<Ticker, QDate, AdjClose>
```

When COIN is used, the receiver can use the data sources as if they were in the receiver context; in this case, the receiver can issue the following query against the simplified schema to obtain adjusted close prices of IBM stock in Frankfurt during December 20, 1998 and January 10, 1999:

Q1:
```
select QDate, AdjClose from YHFrankfurt
where Ticker="IBM.f" and QDate >="12/20/1998" and QDate =<"01/10/1999";
```

This query cannot be executed as is because of the source's inability in evaluating inequalities on "QDate"; even if it could, it does not return meaningful data to the user. Comparing the context definitions for the source and the receiver in Fig. 8, we notice that there are currency and date format differences. The currency assumed in the source also changed within the specified date range. These capability restrictions, semantic differences and the change of semantics are recognized by the COIN mediator, which subsequently generates the following mediated Datalog[7] query:

[7] Datalog is a set-oriented, non-procedural, and function-free logic programming language designed for use as a database language. A Datalog query is the logical equivalent of a SQL query. We use the predicate *answer* to simulate projection; other predicates correspond to relations or selection conditions. For example, a Datalog query for SQL query Q1 is: *answer(QDate,Price):-yhfranfurt("IBM.f",QDate,Price,_,_,_,_,_), QDate>="12/20/1998", QDate=<"01/01/1999"*. A "-" in a predicate represents an unnamed argument of the predicate. Further detail of Datalog can be found in [1].

MDQ1:
```
answer(V6, V5):-
    olsen("DEM", "USD", V4, V3),
    datecvt(V3, "MM/dd/yy", V6, "MM/dd/yyyy"),
    datecvt(V2, "d-MMM-yy", V6, "MM/dd/yyyy"),
    V5 is V1 * V4,
    yhfrankfurt("IBM.f", V2, V1, "Dec", "20", "1998", "Dec", "31", "1998").

answer(V6, V5):-
    olsen("EUR", "USD", V4, V3),
    datecvt(V3, "MM/dd/yy", V6, "MM/dd/yyyy"),
    datecvt(V2, "d-MMM-yy", V6, "MM/dd/yyyy"),
    V5 is V1 * V4,
    yhfrankfurt("IBM.f", V2, V1, "Jan", "1", "1999", "Jan", "10", "1999").
```

The corresponding SQL query generated by the COIN mediator is:

MQ1:
```
select datecvt.date2, (yhfrankfurt.adjClose*olsen.rate)
from   (select 'DEM', 'USD', rate, ratedate from   olsen
        where  exchanged='DEM' and expressed='USD') olsen,
       (select date1, 'MM/dd/yy', date2, 'MM/dd/yyyy' from   datecvt
        where  format1='MM/dd/yy' and format2='MM/dd/yyyy') datecvt,
       (select date1, 'd-MMM-yy', date2, 'MM/dd/yyyy'
        from   datecvt
        where  format1='d-MMM-yy' and format2='MM/dd/yyyy') datecvt2,
       (select 'IBM.f', qDate, adjClose, 'Dec', '20', '1998', 'Dec', '31', '1998'
        from   yhfrankfurt where  Ticker='IBM.f'
        and    StartMonth='Dec' and StartDay='20' and StartYear='1998'
        and    EndMonth='Dec' and EndDay='31' and EndYar='1998') yhfrankfurt
where  datecvt2.date1 = yhfrankfurt.qDate
and    datecvt.date2 = datecvt2.date2 and  olsen.ratedate = datecvt.date1
union
select datecvt3.date2, (yhfrankfurt2.adjClose*olsen2.rate)
from   (select 'EUR', 'USD', rate, ratedate from   olsen
        where  exchanged='EUR' and  expressed='USD') olsen2,
       (select date1, 'MM/dd/yy', date2, 'MM/dd/yyyy' from   datecvt
        where  format1='MM/dd/yy' and  format2='MM/dd/yyyy') datecvt3,
       (select date1, 'd-MMM-yy', date2, 'MM/dd/yyyy' from   datecvt
        where  format1='d-MMM-yy' and  format2='MM/dd/yyyy') datecvt4,
       (select 'IBM.f', qDate, adjClose, 'Jan', '1', '1999', 'Jan', '10', '1999'
        from   yhfrankfurt where  Ticker='IBM.f'
        and    StartMonth='Jan' and StartDay='1' and StartYear='1999'
        and    EndMonth='Jan' and EndDay='10' and EndYear='1999') yhfrankfurt2
where  datecvt4.date1 = yhfrankfurt2.qDate and datecvt3.date2 = datecvt4.date2
and    olsen2.ratedate = datecvt3.date1
```

The SQL syntax is a bit more verbose, so we will examine the more concise Datalog query MDQ1. It has two sub-queries: one for the time period from December 20, 1998 to December 31, 1998, the other for the time period from January 1, 1999 to January 10, 1999. This is because the currency assumed in the source is Deutschmark in the first period and is Euro in the second period, each needing to be processed separately.

Let us focus on the first sub-query for the moment, which is reproduced in Fig. 9 with line numbers and annotations added. Line 6 queries the *yhfrankfurt* source. Notice that the date range has been translated to equalities of the six attributes of *month*, *day*, and *year* of start date and end date of the actual schema; the values for month are now in the format required by the source, i.e., "Dec" for December. Variable V2 corresponds to "QDate", V1 corresponds to "AdjClose". None of them are in line 1 to be reported back to the user; the code in lines 2-5 has the instructions on how to transform them to V6 and V5 as values to be returned to the user.

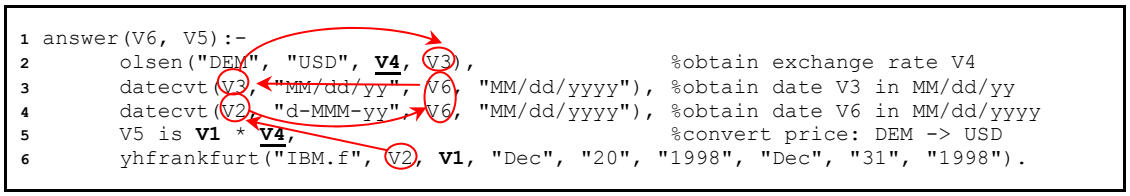


Fig. 9. Reconciliation of semantic differences in MDQ1.

The procedural reading of the code is:

- *line 4* converts "QDate" (V2) from the source format to the format expected by user (V6), i.e., from "d-MMM-yy" format (e.g., 20-Dec-98) to "MM/dd/yyyy" format (e.g, 12/20/1998);
- *line 3* converts V6 (from *line 4*) to V3 so that V3 has the format expected by source *olsen*, i.e., it converts date format from "MM/dd/yyyy" (e.g, 12/20/1998) to "MM/dd/yy" (e.g, 12/20/98);
- *line 2* queries the *olsen* source to obtain exchange rate (V4) between Deutschmark (DEM) and U.S. dollar (USD) for the date given by V3; and
- *line 5* converts "AdjClose" (V1) to USD using the exchange rate (V4) from *line 2*.

The second sub-query is almost the same except that it deals with a different date range within which the currency difference is EUR v. USD instead of DEM v. USD.

When the mediated query is executed, the user receives data instances[8] as shown in the left pane of Fig. 10. For comparison, we also show the "raw" data from the source; notice that the unusual abrupt price drop in the raw data (which is actually due to the change in currencies) no longer appears in the mediated data.

Mediated results

QDate	AdjClose
01/08/1999	91.65
01/07/1999	90.10
01/06/1999	92.94
01/05/1999	88.28
01/04/1999	88.61
12/30/1998	89.27
12/29/1998	90.54
12/28/1998	90.14
12/23/1998	88.06
12/22/1998	84.84
12/21/1998	84.96

(user format) (in USD)

Non-mediated results (original data)

QDate	AdjClose	
8-Jan-99	78.67	EUR
7-Jan-99	77.22	
6-Jan-99	79.15	
5-Jan-99	74.81	
4-Jan-99	75.05	
30-Dec-98	149.13	DEM
29-Dec-98	151.54	
28-Dec-98	151.54	
23-Dec-98	147.2	
22-Dec-98	141.89	
21-Dec-98	141.41	

(original format)

Fig. 10. Mediated and non-mediated data instances.

We have also applied the COIN technology to the other examples. For detailed descriptions, interested readers are referred to [15] for the corporate householding scenario (where we illustrate how entity aggregational ontological heterogeneity is resolved), and [21] for the Yugoslavia example (where we show how temporal entity aggregational heterogeneity is resolved). We are currently extending COIN to address the problem of relationship aggregational ontological heterogeneity.

[8] Mediated results are rounded for easy reading.

4. Concluding Remarks

We are in the midst of exciting times – the opportunities to access and integrate diverse information sources, most especially the enormous number of sources provided over the web, are incredible but the challenges are considerable. It is sometimes said that we now have "more and more information that we know less and less about." This can lead to serious "data quality" problems caused due to improperly understood or used data semantics, as illustrated by the situation described in Fig. 11.

Unit-of-Measure mixup tied to loss of $125 Million Mars Orbiter

"NASA's Mars Climate Orbiter was lost because engineers did not make a simple conversion from English units to metric, an embarrassing lapse that sent the $125 million craft off course … The navigators [JPL] **assumed metric units** of force per second, or newtons. In fact, the numbers **were in pounds** of force per second as supplied by Lockheed Martin [the contractor]."

Source: Kathy Sawyer, *Boston Globe*, October 1, 1999, pg. 1.

Fig 11. Examples of consequences of misunderstood data semantics

The effective use of data semantics and context knowledge is needed to enable us to overcome the challenges described in this paper and more fully realize the opportunities. A particularly interesting aspect of the context mediation approach described is the use of context to describe the expectations of the receiver as well as the semantics assumed by the sources.

In this paper, we identify the kinds of semantic heterogeneities that can cause data quality problems. Then we show how COIN technology can be used to capture context knowledge and improve data quality by automatically reconciling semantic differences between the sources and the receivers. An important aspect of this approach is that COIN is a flexible and scalable technology. As shown in [20], the number of component conversions that need to be specified depends on the number of modifiers in the ontology and the number of unique values of each modifier; it does not depend on the number of sources and receivers involved, *N*. When *N* is large, COIN approach requires one to several orders of magnitude less conversions to be specified than other approaches that hard-code the conversions. This is not surprising because the mediator can be thought of as an automatic code generator – it can generate composite conversions using a small set of component conversions and supply appropriate parameters depending on contexts. Through demonstrations, we have shown that COIN can be used to solve many data quality problems caused by semantic heterogeneities.

We find that the interplay of data quality and data semantics is interesting and has practical significance. This paper presents only some initial work in this area. For future research, we plan to identify other semantic heterogeneities that affect data quality either in the source or from the receiver's perspective. Then we extend COIN-based system to facilitate automatic reconciliation of such heterogeneities. Ultimately, we expect to develop a unifying framework for analyzing data quality from data semantics perspective and applying semantic interoperability technologies to improving data quality.

Acknowledgements

Work reported herein has been supported, in part, by Banco Santander Central Hispano, Citibank, Defense Advanced Research Projects Agency (DARPA), D & B, Fleet Bank, FirstLogic, Merrill Lynch, MITRE Corp., MIT Total Data Quality Management (TDQM)

Program, PricewaterhouseCoopers, Singapore-MIT Alliance (SMA), Suruga Bank, and USAF/Rome Laboratory.

References

[1] S. Ceri, G. Gottlob, and L. Tanca, "What You Always Wanted to Know About Datalog (And Never Dared to Ask)", *IEEE Transactions on Knowledge and Data Engineering,* **1**(1)**,** 146-166, 1989.
[2] P.P. Chen, "The Entity-Relationship Model: Toward a Unified View of Data," ACM Transactions on Database Systems, 1(1) (1976) 1-36.
[3] H. T. El-Khatib, M. H. Williams, L. M. MacKinnon, and D. H. Marwick, "A framework and test-suite for assessing approaches to resolving heterogeneity in distributed databases," Information & Software Technology, vol. 42, pp. 505-515, 2000.
[4] R. Elmasri, J. Weeldreyer, A. Hevner, "The Category Concept: an Extension to the Entity-Relationship Model", Data & Knowledge Engineering, 1(1) (1985) 75-116.
[5] A. Firat, S. E. Madnick, and M. D. Siegel, "The Cameleon Web Wrapper Engine," in: Workshop on Technologies for E-Services (TES'00), Cairo, Egypt, 2000.
[6] A. Firat, S. E. Madnick, and B. Grosof, "Financial Information Integration in the Presence of Equational Ontological Conflicts," in: 12th Workshop on Information Technology and Systems (WITS), Barcelona, Spain, 2002.
[7] A. Firat, "Information Integration using Contextual Knowledge and Ontology Merging," PhD Thesis, MIT, 2003.
[8] T. Frühwirth, "Theory and Practice of Constraint Handling Rules," Journal of Logic Programming, vol. 37, pp. 95-138, 1998.
[9] C. H. Goh, "Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems," Ph.D. Thesis, MIT, 1997.
[10] C. H. Goh, S. Bressan, S. Madnick, and M. Siegel, "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," ACM TOIS, vol. 17, pp. 270-293, 1999.
[11] V. Kashyap and A. P. Sheth, "Semantic and Schematic Similarities Between Database Objects: A Context-Based Approach," VLDB Journal, vol. 5, pp. 276-304, 1996.
[12] A. C. Kakas, A. Michael, and C. Mourlas, "ACLP: Abductive Constraint Logic Programming," Journal of Logic Programming, vol. 44, pp. 129-177, 2000.
[13] M. Kiffer, G. Laussen, and J. Wu, "Logic Foundations of Object-Oriented and Frame-based Languages," J. ACM, vol. 42, pp. 741-843, 1995.
[14] S. E. Madnick and R. Wang, "The Inter-Database Instance Identification Problem in Integrating Autonomous Systems," in: 5th International Conference on Data Engineering (ICDE'89), Los Angeles, CA, 1989.
[15] S. Madnick, R. Wang, and X. Xian, "The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality," Journal of Management Information Systems, vol. 20, pp. 41-69, 2004.
[16] C. F. Naiman and A. M. Ouskel, "A classification of semantic conflicts in heterogeneous database systems," Journal of Organizational Computing, vol. 5, pp. 167-193, 1995.
[17] S. Ram and J. Park, "Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data and Schema-Level Semantic Conflict," IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 189-202, 2004.
[18] A. Rosenthal, L. Seligman, and S. Renner, "From Semantic Integration to Semantics Management: Case Studies and a Way Forward," ACM SIGMOD Record, vol. 33, pp. 44-50, 2004.
[19] Z. Schreiber, "Semantic Information Management (SIM)," Unicorn, White Paper, 2003.
[20] H. Zhu and S. E. Madnick, "Context Interchange as a Scalable Solution to Interoperating Amongst Heterogeneous Dynamic Services," in: 3rd Workshop on eBusiness (WEB), Washington, D.C., 2004.
[21] H. Zhu, S. E. Madnick, and M. D. Siegel, "Representation and Reasoning About Changing Semantics in Heterogeneous Data Sources," in Semantic Web and Databases: Second International Workshop (SWDB 2004), vol. LNCS 3372, C. Bussler, V. Tannen, and I. Fundulaki, Eds., 2005, pp. 127-139.

Stuart Madnick is the John Norris Maguire Professor of Information Technology, Sloan School of Management and Professor of Engineering Systems, School of Engineering at the Massachusetts Institute of Technology. He has been a faculty member at MIT since 1972. He has served as the head of MIT's Information Technologies Group for more than twenty years. He has also been a member of MIT's Laboratory for Computer Science, International Financial Services Research Center, and Center for Information Systems Research. Dr. Madnick is the author or co-author of over 250 books, articles, or reports including the classic textbook, Operating Systems, and the book, The Dynamics of Software Development. His current research interests include connectivity among disparate distributed information systems, database technology, software project management, and the strategic use of information technology. He is presently co-Director of the PROductivity From Information Technology Initiative and co-Heads the Total Data

Quality Management research program. He has been active in industry, as a key designer and developer of projects such as IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system. He has served as a consultant to corporations, such as IBM, AT&T, and Citicorp. He has also been the founder or co-founder of high-tech firms, including Intercomp, Mitrol, and Cambridge Institute for Information Systems, iAggregate.com and currently operates a hotel in the 14th century Langley Castle in England. Dr. Madnick has degrees in Electrical Engineering (B.S. and M.S.), Management (M.S.), and Computer Science (Ph.D.) from MIT. He has been a Visiting Professor at Harvard University, Nanyang Technological University (Singapore), University of Newcastle (England), Technion (Israel), and Victoria University (New Zealand).

Hongwei Zhu is a Research Scientist at the Information Quality Program at MIT. His research interests include the development of technologies to enable meaningful information sharing, and theories to address policy issues related to information sharing/data reuse. He holds a Ph.D. in Technology, Management, and Policy form MIT, where he worked on the Context Interchange Project at the Sloan School of Management. Prior to coming to MIT, he was a software engineer and IT consultant developing web based solutions for both private sector and government agencies.

Example Semantic Web Applications

Introduction

Other lessons address what the Semantic Web is, what are key characteristics of a Semantic Web application, and where its strengths lie compared to most traditional technologies *(coming soon!)*. This lesson presents several specific, successful examples of Semantic Web applications in order to bring these lofty ideas down to reality.

As more case studies come up across the Web, we will try to collect some of the best on this page. Let us know if we are missing one!

Prerequisites

- What Makes a *Good* Semantic Web Application?

Case Studies

- Supply Chain Management—Biogen Idec
- Media Management—BBC
- Data Integration in Oil & Gas—Chevron
- Web Search and Ecommerce

Today's Lesson

When possible, the specific case studies illustrated here pertain to specific corporate projects. Generally speaking, well-known companies are not willing spend money on newer technology unless older, more established techniques either will not work or are outside of their budget for a specific problem. Therefore, these specific corporate use cases tend to highlight applications of Semantic Web technologies that have proven themselves to be of value.

The following high-level summaries include links to further details about each case study discussed.

Supply Chain Management – Biogen Idec

Biogen Idec—a pharmaceutical maker best known for its manufacturing of drugs used to treat multiple sclerosis—manages its global supply chain using Semantic Web technologies. As a class of problems, supply chain management includes many features that make it ripe for applying Semantic Web Technologies, specifically: {C}{C}

- The data being managed changes constantly.
- The required views on those data (e.g., calculations, KPIs, etc.) change constantly.
- A great deal of cross-organizational collaboration takes place, with large volumes of data being conveyed between suppliers at every level of the supply chain.

Furthermore, Biogen's specific industrial requirements make the use of traditional technologies for supply chain management particularly challenging.

- The types of material that a high-tech company such as Biogen Idec ships change over time, and as a result, the properties of these materials are also constantly changing.
- The Key Performance Indicators (KPIs) currently being optimized by high tech companies change very quickly.
- Rules and regulations change, requiring different kinds of data to be captured over time.
- Supply Chain Managers are not IT professionals, so they need to be able to see, understand, and manipulate the data being tracked directly, without having to traverse an additional level of organizational indirection. Keep in mind that the term Semantic in

Subscribe to Semantic University

Further Reading

- Discuss this Lesson in its Forum
- The Semantic Web Has Gone Mainstream! Wanna Bet?—in this article, Juan Sequeda includes an amazing number links to Semantic Web successes to bolster his argument that it has, in fact, gone mainstream.
- W3C-curated list of Semantic Web Case Studies and Use Cases
- Case Studies on SemanticWeb.com

Latest Content

- Semantic Web Design Patterns—Application Patterns
- Semantic Web Design Patterns—Data Publishing Patterns
- Semantic Web Design Patterns—Data Management Patterns
- What is Linked Data?
- What is JSON-LD?
- Semantic Web Design Patterns—Modeling Patterns
- Semantic Web Design Patterns—Identifier Design Patterns
- RDF vs. XML
- SPARQL vs. SQL - Intro
- SPARQL Nuts & Bolts

Semantic Web means that by definition, the data model is transparent to subject matter experts, not only technologists.

- Suppliers change over time and are located in new regions and countries, possibly requiring new language localization, currencies, etc., and often requiring new data connectivity to new third party systems.

Semantic Web technologies give supply chain managers and officers the ability to manage all of this complexity reliably and efficiently.

To read the full case study, see the original article in American Laboratory.

Media Management—BBC

By far the most public usage of Semantic Web technologies is the website for the British Broadcasting Corporation (i.e., the BBC). In 2010, their entire World Cup website was powered by Semantic Web technologies, as was reported on ReadWriteWeb and SemanticWeb.com. Even today, large portions of their public website are run on Semantic Web technologies.
The BBC is not the only media company that is using Semantic Web technologies. Time Inc., Elsevier, and the Library of Congress all also have production systems built using Semantic Web technologies.

The process of storing, sorting, and presenting media has many qualities that benefit from the utilization of Semantic Web technologies:

- Unstructured information.
- Significant cross-document relationships and annotations. Documents have authors, which have written other documents; documents include citations; they have multiple revisions. Managing these relationships using traditional relational databases can get very messy. They in fact do not even attempt to solve this problem, and CMS systems do very poorly at searching on large corpuses.
- Constantly changing usage patterns. Websites have to change to stay fresh in their designs. Links between pages, relationships between videos and pages, links to blogs, etc. will all change over time.

To read the full case study, see the original article at the W3C website. Furthermore, SemanticWeb.com keeps an active list of BBC activity, including links to presentations and press releases related to the Semantic Web.

Data Integration in Oil & Gas—Chevron

For many years, Chevron has been experimenting with Semantic Web technologies in a range of applications.

100 years ago, drilling oil was little more complicated than sticking a pipe in the ground. These days, however, everything from discovery to production is incredibly data intensive. Every day, a single offshore rig will produce terabytes of data containing critical information that can help predict mechanical failures and other anomalies. Every time an error disrupts production on an active rig, costs can soar to tens of millions of dollars a day. Understandably, operators in this field are under an enormous amount of pressure.

Semantic Web technologies enable engineers and researchers to combine arbitrary data in arbitrary ways in an attempt to better understand and predict daily oil field operations. Some of the many high-level considerations that are not handled well by traditional technologies include:

- A lack of well-defined results. By their very nature, many activities throughout the energy industry are experimental. When the end goal state is undefined, it might become necessary to change direction at any point.
- A lack of industry data standards. All data integration is basically ad hoc.
- A massive turnover of technology. Every new device emits new parameters that must be tracked alongside existing data.

To be sure, certain activities in the industry are predictable in a manageable way, but many are not.

The following key business drivers were specifically identified by Chevron (as excerpted directly from the case study):

- "There are a million miles of spaghetti eaten every day!" The same can be said about data in the oil and gas Industry. A large amount of data is generated every day from multiple sources such as seismic measurements, well records, drilling figures, transportation numbers, and marketing statistics. Integrating these heterogeneous data to capitalize on their information value has so far proven to be complex and costly.

their information value has so far proven to be complex and costly.

- These data exist in a structured form in databases, and in semi-structured forms in workbooks and documents such as reports and multimedia collections. To deal with both the flood of information as well as the range of heterogeneous data formats, a new approach was needed for information searching and access.
- For the major capital projects (see application examples below) in the industry, information needs to be standardized and integrated across systems, disciplines, and organizational boundaries. This information integration will enable better decision-making within collaborations, as high-quality data will become more accessible in a timely fashion.

To read the full case study, see the original article on the W3C. Also, a key practitioner from that project, Roger Cutler, gave an exceedingly frank and lucid interview which is well worth a read.

Web Search and Ecommerce

Search engines genuinely benefit from having access to extra metadata in order to return more relevant results. In fact, the biggest players in the industry are investing heavily in standards that encourage companies to annotate their web pages with significantly more structure, which was one of the original intents of the Semantic Web vision in the first place. RDF itself can even be embedded into web pages via RDFa.

Facebook developed the Open Graph Protocol, which is very similar to RDF. Microsoft, Google, and Yahoo use Schema.org, which has an RDFa representation. The Ecommerce Industry has GoodRelations, which also uses RDFa. These frameworks are all now actively being used to bring users a better web experience.

An excellent and specific case study on this usage of Semantic Web technologies is Best Buy. They adopted GoodRelations for their website and saw an unbelievable increase in hits and conversions. Jay Myers has presented at numerous conferences, and his work on the subject can be found all over the web.

About the Author

Rob Gonzalez
Editor-in-Chief, Semantic University, and Director of Product Management & Marketing, Cambridge Semantics
On Twitter: @chirping_gonzo