Using Cyber-enabled Transaction Data to Study Productivity and Innovation in Organizations*

Carol Corrado
The Conference Board

Julia Lane
National Science Foundation

May 2009

* This paper was originally prepared for the conference, *Advancing the Study of Innovation and Globalization*, held in Nuremberg, Germany, May 29-30, 2009. It integrates and extends work that was done for the National Science Foundation (Corrado 2008) and the German Research Foundation (Lane 2009), which we thank for support.

Contact information for authors:
Corrado: carol.corrado@conference-board.org
Lane: jlane@nsf.gov

1. Introduction

Innovation is recognized as being a driving force contributing to United States competitiveness. Yet the current empirical understanding of the innovation ecosystem is insufficient to guide decision makers. A growing understanding of innovation and new cyber-enabled capacities to collect and integrate data about individuals and organizations offer expanded potential for scientists, policy-makers and organizations themselves to understand the way in which innovation contributes to key national priorities such as the generation of new knowledge, and the creation of new jobs, income and wealth. New cyber-enabled advances in confidentiality protection also now make possible the analysis of sensitive data without revealing individual identities – so that researchers can generalize and replicate scientific results.

The capture of data about innovation *within* organizations could be advanced by using sensors, radio frequency identification RFID) chips, videos, cell phones and GPS which provide new ways to capture information about humans and about the way they interact with each other (Lane, 2009). Just as Deming worked with Japanese and US business to study and understand the production of physical goods, social and computer scientists could work with businesses to study successfully innovative project teams.

The capture of data about innovation *across* organizations can be advanced by using new cyber tools that both scrape the web and make sense of very heterogeneous sources. The field of business intelligence recognizes that the World Wide Web provides a rich source of data on corporate America,1 especially when used in combination with internal knowledge such as digitized ledgers and activity-based cost systems. Further, by identifying common metrics for innovation processes within organizations (or the proto-organizations within which innovation occurs) and linking that microdata to a longitudinal resource and infrastructure capable of sustaining the scientific inquiry, knowledge needed by decision makers will emerge (Corrado 2008).

The scientific basis for studying data upon which a nation's innovation policy may be based must rely on widespread *access by researchers*. Such access is central to ensuring that the work is generalizable and replicable. Social scientists need to work with computer scientists to ensure that the research data are accessible (possibly by establishing virtual organizations); the scientists also need to work to see that the *privacy and confidentiality* issues associated with studies based on microdata from deep within organizations are addressed.

A nascent community has begun to think about these issues and how to address the *questions and data gaps* in our current understanding of innovation. A recent workshop[2] that brought social scientists together with computer scientists and businesses found that "the creation and analysis of representative information are core elements of the scientific endeavor. No less fundamental is the need to replicate analysis and protect respondent identity. For a variety of reasons, currently available business microdata generally do not jointly meet these criteria of

[1] The research of Chen, Chau, and Zeng (2001) confirms that companies seeking a competitive advantage in the marketplace find they must use a combination of internal knowledge and external sources, such as the Web.
[2] http://www.conference-board.org/events/nsf/Workshop_Report_Final.pdf

scientific inquiry (representative coverage, researcher access/replicability, and confidentiality protection). Developing the needed innovation business microdata and research along with an infrastructure for access and protection requires solving a series of technical and social challenges." It calls for a national research infrastructure for the study of organizations.

This paper reviews how cyber tools can be used to capture and advance the creation of new data for the study of organizations and innovation and what elements are needed for a data research infrastructure to meet the criteria of scientific inquiry. The paper then reviews important questions and data gaps, as identified by a group of social scientists, for researchers to address in order to advance our understanding of the innovation process and the science of innovation policy.

2. Key issues in developing a database on organizations

An important challenge facing social scientists is how to *measure innovation in a business context* and develop a broad understanding of innovation's processes, lifecycles, and role in the economy and global business environment. Innovation can be seen as a process whereby organizations put something new (research results, ideas, designs, employee knowledge) to commercial use or financial gain.[3] The study of innovation must involve a unit (or units) of observation applicable both *within and across* organizations, and preferably scalable. Data on innovation inputs and the *business outcomes* expected and/or actualized from those inputs are the major gaps in our current data system.

[3] This view is consistent with the definition used in the report to the Secretary of Commerce by the Advisory Committee on Measuring Innovation in the 21st Century Economy.

What are the key issues? A number of areas were identified at the workshop

Unit of analysis

What is the most fruitful level of analysis for new scientific research on innovation? One approach is to have a *project-based unit of analysis* within a given business organization (or proto-organization where innovation occurs). The project is the basic unit of production in many services firms and the basic "unit of innovation" in many others. Although data on innovation projects may be obtained from company records for certain organizations, determining and developing the "unit of innovation" itself is an important subject of research for many others. Characteristics of this unit of innovation include the following:

- The unit should capture the entire lifecycle of an innovation (or the expected lifecycle).
- Depending on the precise research question, the unit should be scalable upwards (firms, groups or networks of firms) or downwards (teams, social networks, entrepreneurs).
- The unit must be associated with an outcome that determines the degree of success of the research project/initiative/idea.

A project-level unit of analysis provides natural "scope advantages". In particular, conducting innovation research at the project-level captures the development of customized services and creative solutions to general problems, areas beyond the scope of existing studies whose focus is scientific R&D yet especially relevant to business strategists (understanding the complete value chain) and economic analysts (understanding the service economy in the United States). The project-based unit of analysis is especially relevant for cutting across the multiple organizations (alliances, universities) that play important roles in innovations with long lifecycles and whose processes are complex (e.g., "open" innovation processes).

Owing to the multinational nature of many businesses, global considerations increasingly enter business strategic decisions, and innovative activity of U.S.-headquartered firms is not necessarily located in the United States. As a result, many research questions need data relevant for the multinational, global domain.

- The project-level unit of analysis is amenable to the collection of innovation data across and within national boundaries for multinational firms.
- In work in this area, the results of the "community innovation surveys" conducted in many OECD countries, as well as BEA's surveys of multinational companies, and other sources of global business information, were considered important complements of the new infrastructure.

Innovation process microdata

What data are needed to determine the economic and social value created by innovation in organizations? What are the characteristics of successful innovations?

- To determine the economic value created through innovation, data on (1) the full costs of an innovation project over its lifecycle and (2) a measure or measures of the outcome of the project (preferably one in a dollar metric) are needed.

- In general, detailed data on workers—their skills, their responsibilities, and their knowledge—including their flows across companies were desired for transformative research on the combined process of entrepreneurship and innovation.
- Data on the social and cultural aspects/determinants of innovation were also desired, especially for exploring the emerging area of social networks.

The creation of basic data on *innovation net outcomes* allows the study of the determinants and characteristics of innovation successes and failures. The roles of: organizational practices (employment and management); organizational characteristics (employee knowledge and skills, business model, IT use); environmental and cultural factors (location and networks); entrepreneurial factors (firm age and origin); as well as other factors (dynamics) can be examined in terms of degree of success.

Existing studies have associated many of the above-mentioned factors with firm-level market valuations and/or labor productivity. But the established associations generally are not structural. How *do* firms appropriate the knowledge of their employees? How *do* enhancements the work environment promote innovation? Creating a new data infrastructure opens richer and deeper opportunities for exploring these questions.

Much of the economic and social data called for are relatively basic because, as previously noted, some companies currently keep records of costs and margins along project lines. For other companies, the underlying production and innovation processes in participating companies will need to be identified and the project-level data on units of innovation created accordingly. And, where innovations are the outcome of a diffuse, creative and risky process with a long time lag between spending and payoff, specifying and determining the basic data and the role of expectations will be challenging.

Business function microdata

Research data on innovation processes from selected companies are unlikely to be representative of innovation inputs and outcomes in the economy as a whole. What can be done to preserve the scientific inquiry?

The collection of representative data by business function/process is a necessary component of a data infrastructure for the study of innovation and organizations. Business function concepts have been found to have substantial meaning and applicability for respondents to business surveys, largely because the concepts are grounded in the popular value chain model of firm activities introduced by Michael Porter in his 1985 best-selling book, *Competitive Advantage*.

Business processes and business functions include procurement, operations, products and services development, and the like. The approach would be to sample the universe of employer firms and collect selected economic data (e.g., total spending and employment costs) by business function/process. The Mass Layoff Statistics program of the Bureau of Labor Statistics has experimented with the collection of data by business function and finds that most establishments define their activities in terms of business functions (Brown, 2008).[4]

---

[4] See also Sturgeon *et al*. (2006) and Lewin *et al*. (2008).

The linking of innovation project-level data with business function/process-level data would yield a new microdata laboratory for studying innovation and organizations. The inclusion of the business function/process-level survey data uniquely provides:

- A rich longitudinal resource for standardizing and benchmarking the data and findings from the project-level innovation process research and for linking them to other sources of data useful for studying innovation.

- The ability to design new innovation indicators; for example, national totals of business spending on “new product and/or service development” and “strategic management/business process development” are an indicator of business investments in innovation.

This research data infrastructure—the base microdata laboratory and ongoing survey apparatus—could be supported and managed in a fashion similar to the General Social Survey: Core statistics collected every one or two years; supplemental data modules added to address specific research questions and business and policy issues of the day; and wide researcher access via a system with appropriate safeguards and standards.

<u>The contribution of innovation to the national economy</u>

What does innovation do for the national economy? What are the indicators of future innovation success and/or failure? How can gains in social welfare be fostered through innovative activity?

Because broad indicators of innovative activity are next to nonexistent, policy analysts still rely on science and engineering indicators—data on patents, R&D inputs, the S&E workforce, and the like as gauges of inventive/innovative activity. Activities of the modern business organization, such as market research, "soft" design and development, the creation of entertainment and artistic originals, and investments firms make in training employees and developing new business models and strategies—activities associated with innovation—are unmeasured and missed in the discourse.

Measures of innovation inputs, however broad, are limited as indicators because the productivity of the inputs themselves generally is not known. This long has been a limitation of the practice of using S&E inputs to look for areas of under-investment to suggest how policy-makers should allocate resources to promote economic growth. Addressing the issues discussed above would yield new insights, stronger empirics, and thereby a strengthened understanding of the role of innovation in economic growth. For example:

- Basic information on innovation projects (lifetimes and relative costs/prices, for example) can be used to improve the placing of different types of national-level innovation investments on the same footing.
- The ability to use the new microdata to further study the connections between commercial success and government-sponsored research and entrepreneurship helps policy-makers formulate strategies for advancing the rate and direction innovative activity.

In sum, a dataset on organizations could be used to address four key sets of questions.

The first of these is to develop a scientific way of measuring *what* innovation is: identifying the units, the scales, and the level and trajectory of activity. This is the first step to determining how economic value gets created through innovation. This would include such factors as determining the rate of return to projects within a company by developing ways of measuring the full cost of the inputs to innovation (over its complete lifecycle) and the business outcomes expected and/or actualized from those inputs was a priority. Such an approach would permit businesses to better understand the distribution of rates of return for different projects, including appropriate time horizons.

The second is to advance an understanding of *how* and *why* innovation takes place. This is necessary to identify the inputs to innovation (including knowledge itself), the determinants of successful innovations, and the factors that affect how innovations diffuse, such as social networks and geography. Characterizing the features and practices of organizations, individuals, industries, markets, and nations—and the links among them—that promote innovation (including the skill/talent/training of the workforce) is necessary.

The third is to understand the *consequences* of innovation. This is particularly true in terms of understanding the impact of innovation on aggregate economic activity, but also the effect of outsourcing of parts of value chain in terms of economic vulnerability of locations, unemployment, the associated political outcomes, and corporate social responsibility.

The fourth is to understand the broader *environment*: globalization, technological change, and innovation are interdependent processes in our economy. New ways of communicating exemplified by Web 2.0 will change business's customers, suppliers and partners. Trends in emerging markets and competitiveness will determine the pattern of global engagement. A better understanding of the impact of formal and informal interactions on the boundaries of companies and industries, the diffusion of technology and ideas, and the larger process of value creation can inform business strategy and policy-making.

3. Taking Stock

New needs and previous approaches

Existing data infrastructures are not sufficient for researchers to model, measure, and study the evolving mechanisms whereby innovating enterprises and entrepreneurs create economic value. The call for better data and metrics on innovation was made clear by the America COMPETES Act, the Secretary of Commerce's Advisory Committee report, and the National Academies' report on Understanding Business Dynamics (a panel of the Committee on National Statistics).[5]

Business activity is the basic engine of innovation and economic growth, creating jobs and generating income. Although a large empirical literature has yielded insights into topics that fundamentally affect the business environment (such as taxation, regulation, and technical change), the underlying mechanisms that generate entrepreneurship and foster the innovation process within organizations are not well understood. Until innovation and entrepreneurship

[5] See Haltiwanger, *et al.* (2007) for the Academies' report.

are better measured, modeled, and studied from both within and outside of business organizations—and a more or less commonly accepted body of scientific knowledge emerges—policy formulation, business attitudes, and academic research will remain disconnected.

Several approaches have been taken to create business datasets that researchers can use to increase the scientific understanding about innovation and organizational change. One approach was a partnership between academics and businesses that developed a business database called the PIMS project (Profit Impact of Marketing Strategy). This project created a large panel dataset of firms and provided new insights into business decisions such as market entry, pricing and product quality. This project fell into disuse for a variety of reasons, however, and little academic research has used the data in recent years. Nonetheless, the PIMS project is an example of applied research that pushed the frontiers of business strategy formulation.[6]

Another approach, partially supported by the National Science Foundation, is to provide access to the Census Bureau's Business Register by permitting researchers to work with the data at eight Research Data Centers. The resulting research has generated new insights into firm behavior, job creation and job destruction. A related infrastructure project was the Longitudinal Employer-Household Dynamics (LEHD) program which provided, for the first time, an infrastructure that could analyze the impact of economic turbulence on worker job ladders, career paths and firm performance. These data are not widely used, however, as access costs several thousand dollars a month, the process of proposal approval is arduous, and researchers must travel to one of the eight Data Center sites.

[6] See papers and analysis in Farris and Moore (2004) for further information.

Other approaches have turned to commercial datasets, such as Standard & Poor's COMPUSTAT and the files made available by the Center for Research on Security Prices (CRSP) or Wharton Research Data Services. The availability of these files, which provide financial and accounting information on publicly traded companies, has had a major influence on financial and accounting research. Similarly, datasets like Dunn and Bradstreet and ABI/Inform are often used as sample frames for nongovernmental surveys. Obtaining representative and relevant research data from commercial sources is difficult, however. COMPUSTAT and CRSP are able to cover publicly-held companies only, and the content is largely aimed at serving institutional investors; the Dunn and Bradstreet and ABI/Inform datasets are primarily for marketing purposes. As a result, the use of these datasets for supporting a broad research agenda is highly questionable.

The international community has approached the lack of statistical information on innovation (beyond existing R&D and patent indicators) by developing the Community Innovation Survey (CIS). The CIS has been collected widely throughout Europe and other countries, such as Australia and Japan, since the early 1990s. Though the CIS has become institutionalized, evidence based on CIS surveys has yet to significantly influence the development of policy (in any country with a CIS, much less the United States, which does not have a CIS). The available scientific findings to date are apparently somewhat limited, and they and the available CIS metrics have not been oriented to answering questions that are relevant to policy formulation (Arundel, 2006).

Potential for change

Without an investment in data on organizations that engages (1) use by the best researchers, and (2) the participation of business, while providing the necessary safeguards and researcher access, the present situation is unlikely to change.

The potential for getting additional, needed information through federal statistical agencies is very small: New surveys, even new questions and modules on existing, representative surveys can take up to a decade. Capturing high quality information about key measures, such as technology and personnel practices, is difficult both because of respondent burden and problems with identifying the right respondent. In addition, data collected by federal statistical agencies are often not well suited for amendment because they are collected for “core” statistical purposes — the Census Bureau’s business data collection is primarily structured for our national accounts while the Bureau of Labor Statistics’ data programs are designed to provide information about labor markets and prices.

The call for better data and metrics in the Commerce Secretary’s report on Measuring Innovation in the 21st Century included goals for the Bureau of Economic Analysis (BEA) to incorporate better measures of the inputs to, and output of, innovation in the national accounts. These actions will partially be met when the agency includes scientific R&D as investment in the national accounts in 2012, a significant step, but less than called for in the report. The BEA indicated inadequate data as the reason for not capitalizing investments in innovation beyond scientific R&D (Aizcorbe, et. al. 2009), and we know of no plans for new or expanded government surveys to fill the gap. A multi-year effort by the NSF’s Statistics Resources Division to update their R&D survey is expected to lead to significant improvements in the available data on R&D, however.

The potential for researchers to play a key role in developing an expanded empirical platform for the scientific study of innovation within and across organizations and thereby respond to the innovation knowledge gap is therefore significant. This potential has opened up as new cyber tools and advances in confidentiality are transforming the way in which data can be collected within businesses and as researchers learn more about the nature and empirics of innovation.

Data captured via the World Wide Web generally are complex and unstructured. Because human ability to comprehend the huge amounts of information on the Web is limited, leveraging web-based data requires a combination of human and machine resources to extract knowledge. Machine processing systems have enjoyed considerable success in "mining" web-based data, looking for patterns or other pieces of information to present to human analysts.[7] The potential for collaboration among social and computer scientists in using existing natural language processing and information retrieval technologies to "make sense" of complex unstructured data to study innovation processes is substantial.

[7] See, for example, http://www.filtrbox.com/

Advances in the empirics of innovation are substantial and too numerous to summarize but are notable in that, even as we learn that the innovation process has become increasingly open and global, studies increasingly examine processes and practices *within* organizations (IT use, management practices) and investigate the role of organizational characteristics (location, firm age, HR practices) on productivity and innovation.

Alex Pentland, a professor at the Media Lab at the Massachusetts Institute of Technology who is leading the dormitory research project, was a co-founder of Sense Networks. He is part of a new generation of researchers who have relatively effortless access to data that in the past was either painstakingly assembled by hand or acquired from questionnaires or interviews that relied on the memories and honesty of the subjects.

The Media Lab researchers have worked with Hitachi Data Systems, the Japanese technology company, to use some of the lab's technologies to improve businesses' efficiency. For example, by equipping employees with sensor badges that generate the same kinds of data provided by the students' smartphones, the researchers determined that face-to-face communication was far more important to an organization's work than was generally believed.

Productivity improved 30 percent with an incremental increase in face-to-face communication, Dr. Pentland said. The results were so promising that Hitachi has established a consulting business that overhauls organizations via the researchers' techniques.

Dr. Pentland calls his research "reality mining" to differentiate it from an earlier generation of data mining conducted through more traditional methods.

You're Leaving a Digital Trail. What About Privacy? John Markoff, New York Times, Nov 30, 2008

The ideal data creation/collection effort and infrastructure therefore links a base researcher-driven microdata laboratory with available or newly created longitudinal data and indicators that span the reach of the national innovation ecosystem.

4. Creating Cyber-enabled Data

Within Organizations

The potential to describe minute-by-minute human interactions with the physical environment became reality with the development of RFIDs (radio frequency identification devices) and video technologies. RFIDs can be produced for pennies a unit and emit a wireless signal that enables the bearer to be tracked. Businesses now use the technology routinely to track employees (e.g. to ensure that night guards do their assorted tours at the assorted times) and to track their customer behavior.[4] The potential for social science research is clear – ranging from tracking time use information in a far more granular fashion than from survey data, to the environmental impacts on social behavior to measuring the number and quality of human interactions.

In fact similar technologies are already being used for research purposes to great advantage. In addition to the example used in the inset box, Schunn (2008) uses video data collected from a recent highly successful case of science and engineering, the Mars Exploration Rover, to study the way in which human interactions contributed to the success of the project. While the project both wildly exceeded engineering requirements for the mission and produced many important scientific discoveries, not all days of the mission were equally successful. Schunn uses the video records to trace the path from the structure of different subgroups (such as having formal roles and diversity of knowledge in the subgroups) to the occurrence of different social processes (such as task conflict, breadth of participation, communication norms, and shared mental models) to the occurrence of different cognitive processes (such as analogy, information search, and evaluation) and finally to outcomes (such as new methods for rover control and new hypotheses regarding the nature of Mars).

Of course, human behavior is increasingly captured through transactions on the internet. For example, most businesses, as well as registering with the tax authority, also create a website. It is now entirely possible to use web-scraping technologies to capture up to date information on what businesses are doing, rather than relying on administrative records and survey information. Historical records on businesses can also be created by delving into the repository of web pages on the Wayback Machine (see Figure 1 for an example of the web pages for Citibank). This archive takes snapshots of the web every two months and stores them in the manner shown, providing a rich archive of hundreds of billions of web pages. Individual as well as business behavior can be studied using this archive. Indeed, major NSF grants, such as the Cornell Cybertools award[5], have funded the study of social and information networks using these very large semi structured datasets.

Figure 1: The Wayback Machine: http://www.archive.org/index.php

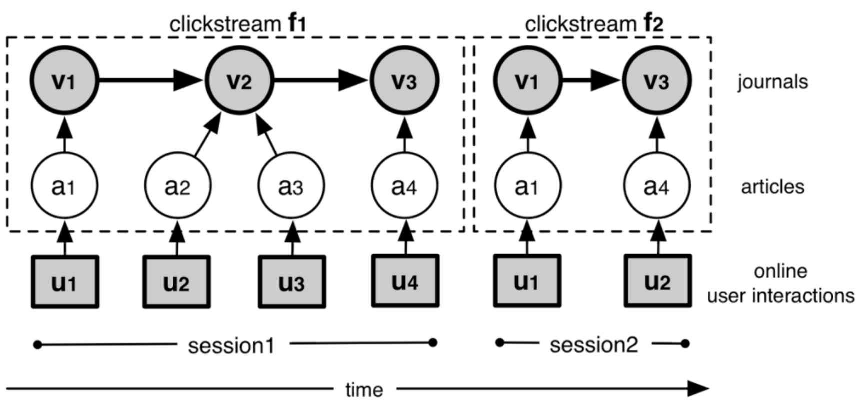


Figure 2

Other ways of collecting information on human behavior from the web include capturing click streams from usage statistics. The MESUR project, for example, has created a semantic model of the ways in which scholars communicate based on creating a set of relational and semantic web databases from over one billion usage events and over ten billion semantic statements.[8] The combination of usage, citation and bibliographic data (see Figure 2) can be used to develop metrics of scholarly impact that go well beyond the standard bibliometric approaches used by academics (Bollen et al. 2009).

Or, in another example following the discussion above, new data could be collected on innovation processes at a project-team level to study factors that generate competitive advantage within firms. The research challenge to social scientists, of course, will be to develop theoretically driven micro-level measures of innovation within organizations. The research challenge to computer scientists is to create cyber-enabled ways for teams to communicate and innovate and capture that information in a structured form, as well as develop ways of capturing data on the process of creativity and insight.

A related but separate research challenge is how to capture data on the role of IT and innovation within organizations, both examining the role of IT as a process enabling innovation and IT as a disruptive technology.

[8] MESUR: Metrics from Scholarly Usage of Resources http://www.mesur.org/MESUR.html

Across Organizations

Capturing detailed high quality information on all firms using existing approaches is too costly and slow to be practical. However, the advent of the internet has not only made vast amounts of new data on firm available, but also created new technologies for harvesting such data. Existing approaches could be combined with new technology to create a *new open source dataset* on business dynamics. Six steps are necessary.

1. *Create a business sample frame*

The first step is to create a file that comprises, as nearly as possible, a universe register of U.S. business organizations. This could be done in at least two ways. One way would be to use IRS data as a sample frame, in the same way as the Survey of Consumer Finances uses IRS data. The frame could be subset to over sample firms in highly innovative industries or with innovative characteristics (Greenia, Husbands-Fealing, and Lane, 2008).

Another approach is to merge a variety of both publicly available and commercial datasets. Publicly available datasets include SEC filings, filings of 5500 forms. Commercial datasets include Compustat, Dunn and Bradstreet and ABI/Inform. The core of the combined dataset would include data elements such as business name (s); address; parent/subsidiary information, industry, sales and employment. Other elements would be added as available. Although this methodology will not capture all small businesses or private equity, appropriate sales and employment weights could be derived for small business by the use of the County Business Patterns data available at the U.S. Census Bureau and possibly inferred for private equity from the research by Davis, Haltiwanger, and Learner (2009).

2. Create a taxonomy

The second step is to create a set of indices that can be used to organize vast amounts of data and create repositories of information. Obvious initial keys include industry, location, and size. Initial keywords might include such terms as "innovation", "technology", or others identified by social scientists.

3. Harvest the web

The third step is to adapt existing web crawlers, such as CiteSeer, to this application. Just as CiteSeer uses academic articles as a sample frame, the proposed application would use the business sample frame described in step 1. The web crawler would "scrape the web" for every mention of every business in the frame 24 hours a day, 7 days a week.[9] Storing, indexing, archiving and curating this vast amount of data is a nontrivial challenge. The expertise of computational linguists would be required to analyze the resulting text and archive it according to the initial taxonomies identified in step 2.[10] Computer scientists would need to be engaged to find efficient ways of storing the information. The database is likely to have both qualitative and quantitative components. This would require the adaption of advanced data-mining tools that have already been developed to capture and quantify a wide variety of text and video information. Successful examples abound: for example, personnel information from business websites is routinely captured and quantified.

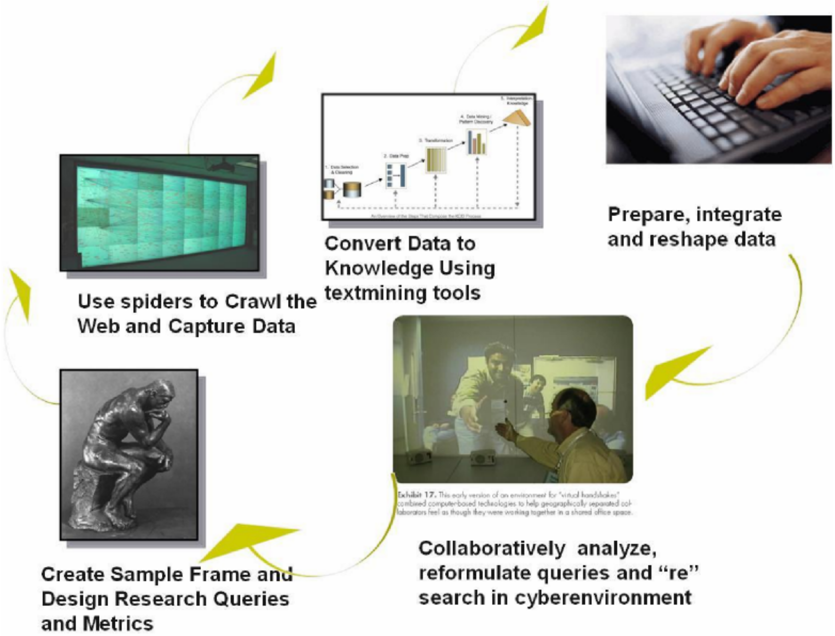
4. Refine the taxonomy

---

[9] A good prototype is the Web laboratory at Cornell University http://www.cs.cornell.edu/wya/weblab/index.html

[10] An example of this is the scientific literature digital library CiteSeer http://citeseer.ist.psu.edu/

Because the resulting database consists of publicly available information, the initial taxonomy can be refined by the user community – which would include businesses, policy makers and academics. Very successful models already exist for this. Although the most visible is Wikipedia, academic examples like the archaeological community's ontologies developed to classify pottery shards at Arizona State University.[11]



5. *Develop quality measures*

A major challenge is developing quality metrics, since the web is a vast unregulated environment. Some sites might initially be assigned high metrics (such as the source business site and sites known to be reputable, such as national media, academic and government sites). Others would be automatically and constantly rated on the quality of the information by means of benchmarking with other sources. Users might develop their own metrics (as is done in commercial ventures like Amazon.com or Ebay).

6. *Creating the data infrastructure*

The vast amount of information collected by means of harvesting the internet will provide a rich contextual backbone. A standard core could be created for all businesses, together with

[11] http://cadi.asu.edu/

longitudinal links.[12] Many models would be possible, for example, a public use core that was available in real time, with user defined data elements and constituencies. The standard core could also be added to the user community (after vetting).

Industry specific modules could be developed in connection with the business community and with Sloan Industry Center researchers. Businesses that agreed to participate in surveys or case studies would receive their summary information together with industry and national benchmarks – potentially in real time. Businesses could control the degree of confidentiality or researcher access, recognizing that greater access would result in more analytical work on their business.

The data could be statistically matched with federal statistical business registers, and do detailed analysis behind federal statistical firewalls.

5. Privacy and Confidentiality Issues

An important challenge to collecting such data is the ethical issues raised by the new capacities to collect data on human beings, particularly a focus on the privacy and confidentiality issues raised by collecting data on the interaction of human subjects.

Also of interest is how to convey the quality of such confidentiality measures to the humans who are the subject of study. Social scientists could expand their current interest in confidentiality to develop approaches that ensure the collaboration and engagement of individuals and

[12] See, for example, the personnel information captured by the Mayflower Group www.mayflower.com

organizations in providing data to the research community, as well as permit the data to be shared so that empirical analyses can be generalized and replicated.

It is worth noting that there is increasing interest by computer scientists in ways in protecting confidentiality so that sensitive data can be collected and analyzed without revealing individual identities – and so that researchers can generalize and replicate scientific results.[13] This interest includes policies for the anonymization and sanitization of the data, retention and storage protocols, transformation prior to dissemination and retaining usability.[14]

6. Analyzing Data

Of course, together with new data, new analytical techniques and new modes of analysis need to be developed. Standard regression analysis and tabular presentations are often inadequate representations of the complexity of the underlying data generation function. There are a variety of reasons for this inadequacy. First, the units of analysis are often amorphous – social networks rather than individuals, firm ecosystems rather than establishments. Second, the structural relationships are typically highly nonlinear, with multiple feedback loops. Third, theory has not developed sufficiently to describe the underlying structural relationships, so "making sense" of the vast amounts of data is a substantive challenge. There has been substantial effort invested in developing new models and tools to address the challenge, however. For example, since a major national priority is to understand the formation and evolution of terrorist networks through the internet and other communication channels, substantial resources have been devoted to the field of

[13] http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5033268&org=CNS
[14] http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf09036

visual analytics. Their research agenda aligns very closely with a potential research agenda for social scientists, focusing as it does on the science of analytical reasoning, visual representations and interaction techniques, data representations and transformations, as well as the production, presentation and dissemination of complex relationships (Thomas and Cook, 2005). It is also worth noting that new partnerships are being formed to address the nontrivial computing challenges.[15]

New cyber-tools also provide an opportunity for social scientists to develop new modes of analysis, such as virtual organizations that study social science data.[16] The opportunity is clear from the way in which ubiquitous information technologies has transformed many facets of human interaction and organization. Tools such as the Grid, MySpace, and Second Life have changed how people congregate, collaborate, and communicate. Increasingly, people operate within groups that are distributed in space and in time that are augmented with computational agents such as simulations, databases, and analytic services which interact with human participants and are integral to the operation of the organization.

Establishing a virtual organization approach would provided the social science community with the ability to move away from individual, or artisan, science, towards the more generally accepted community based approach adopted by the physical and biological sciences. It would provide the community with a chance to combine knowledge about data (through metadata

[15] http://www.nsf.gov/news/news_summ.jsp?cntn_id=111470

[16] This is a group of individuals whose members and resources may be dispersed geographically, but who function as a coherent unit through the use of cyber infrastructure. A virtual organization is typically supported by, and provides shared and often real-time access to, centralized or distributed resources, such as community-specific tools, applications, data, and sensors, and experimental operations.

documentation), augment the data infrastructure (through adding data), deepen knowledge (through wikis, blogs and discussion groups) and build a community of practice (through information sharing).

This opportunity to transform social science through such a organizational infrastructure could potentially be as far-reaching as the changes that have taken place in the biological and astronomical sciences. It is, however, an open research question for the social science data community as to how such an organization should be established: whether the approach should be centralized (like the UK's JISC) or decentralized (like the U.S. National Science Foundation's approach). Similarly, it is an open research question as to the appropriate metrics of success, and the best incentives to put in place to achieve success. However a recent solicitation[17] as well as the highlighting of the importance of the topic in NSF's vision statement,[18] suggests that there is substantial opportunity for social science researchers to investigate these research issues.

7. Concluding thought

In modern economies, economic value is derived increasingly through making and selling ideas. At one time, the production and trade of food was the primary basis of economic value creation, and social and economic thought was grounded in the world of agriculture. The Industrial Revolution created a new social and economic infrastructure: Human beings could not add value

[17] www.nsf.gov/pubs/2008/nsf08550/nsf08550.htm

[18] NSF Cyberinfrastructure Vision for 21st Century Discovery, March 2007

by making and selling things other than food. As a result new theories and new data on manufacturing firms and workers emerged.[19]

The scientific challenge of today is to advance our understanding of how economic value is created through innovation and knowledge appropriation. New data on innovation and knowledge appropriation are needed to represent modern business activity and to guide policy makers in the 21st century economy. Cyber-enabled transaction data grounded in theoretically driven micro-level measures of innovation within organizations offer expanded potential for scientists to meet these needs.

[19] The ferment is well described in Heilbroner's "The Worldly Philosophers" (Heilbroner 1995).

References

Advisory Committee on Measuring Innovation in the 21st Century Economy, U.S. Department of Commerce (2008). Innovation Measurement: Tracking the State of Innovation in the American Economy. Report to the Secretary of Commerce transmitted January 18, 2008. http://www.innovationmetrics.gov/Innovation%20Measurement%2001-08.pdf

Aizcorbe, Ana M., Carol E. Moylan, and Carol A. Robbins (2009). "Toward Better Measurement of Innovation and Intangibles." *Survey of Current Business* 89 (1), 10-23. Available at http://www.bea.gov/scb/pdf/2009/01%20January/0109_innovation.pdf

Arundel, Anthony (2006). "Innovation Survey Indicators: Any progress since 1996?" available at http://www.oecd.org/dataoecd/24/28/37436234.pdf

Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, Chute R, et al. (2009) Clickstream Data Yields High-Resolution Maps of Science. PLoS ONE 4(3): e4803. doi:10.1371/journal.pone.0004803 .

Brown, Sharon P. (2008). "Business Processes and Functions: A New Way of Looking at Employment," *Monthly Labor Review* 131 (12), 51-70. Available at http://www.bls.gov/opub/mlr/2008/12/art3full.pdf

Christensen, Clayton M. (1997). *The Innovator's Dilemma*. Boston: Harvard Business School Press.

Chen, H., M. Chau, and D. Zeng (2001). "CI Spider: a tool for competitive intelligence on the Web." *Decision Support Systems* 34(1), 1-17.

Cockburn, Ian M., Bronwyn H. Hall, and Manuel Trajtenberg, "National Bureau of Economic Research Patent Database: Data Overview," paper presented at the *2007 Kauffman Symposium on Entrepreneurship and Innovation Data*.

Contractor, Noshir (2008). "CI-KNOW: A Tool for Understanding and Enabling the Transformative Power of Cyberinfrastructure in Virtual Communities." presentation at the National Science Foundation (September).

Corrado, Carol A. (2008). Workshop on Developing a New Data Infrastructure for the Study of Organizations and Innovation. Available at http://www.conferenceboard.org/events/nsf/

Duke, Charles (2005). "Remarks" in *Research and Development Data Needs: Proceedings of a Workshop*, p. 17-19. Report of the Board on Science, Technology, and Economic Policy. Washington, D.C.: The National Academies Press, 2005.

Farris, Paul W. and Michael J. Moore, eds. (2004). *The Profit Impact of Marketing Strategy Project: Retrospect and Prospects*. Cambridge, U.K: Cambridge University Press.

Greenia, Nicholas, Kaye Husbands-Fealing, and Julia Lane (2008). "Studying Innovation in Businesses: New Research Possibilities." Paper presented at the CRIW workshop at the 2008 NBER Summer Institute (July). Cambridge, Massachusetts. Available at http://www.nber.org/~confer/2008/si2008/PRCR/lane.pdf

Haltiwanger, John, Lisa M. Lynch, and Christopher Mackie, eds. (2007). *An Integrated Data System for America's Future*. Washington, D.C.: The National Academies Press.

Heilbroner, Robert (1995, 7th edition). *The Worldly Philosphers: The lives, times, and ideas of the great economic thinkers.* New York: Touchstone Press.

Helfat, C.E. and Steven Klepper, "Firm and Industry Evolution and Entrepreneurship (FIVE Project): Data Overview." 2007 Kauffman Symposium on Entrepreneurship and Innovation Data. Available at: http://papers.ssrn.com/paper=1028022

Lane, Julia (2009). "Administrative Transaction Data." In G. Wagner (Ed.), *Improvements and Future Challenges for the Research Infrastructure.* Berlin, Germany: RatsWD.

Lewin, Arie Y., Sivlia Massini, and Carine Peeters (2008). "Why Are Companies Offshoring Innovation? The Emerging Global Race for Talent." *Journal of International Business Studies* (forthcoming). Available at http://faculty.fuqua.duke.edu/ciber/site2006/Downloads/JIBS%20paper.%20active%20voice.%20%20July.14.208.TM%20and%20AYL-3.W-2003.Aug.5.2008.pdf

Maraut, S. *et al*., "The OECD REGPAT Database: A Presentation," OECD Science, Technology and Industry Working Papers, 2008/2, OECD Publishing. doi:10.1787/241437144144

Patton, D. and M. Kenney, "Initial Public Offering Database: Presentation." 2007 Kauffman Symposium on Entrepreneurship and Innovation Data. Available at: http://ssrn.com/abstract=1028004

Porter, Michael (1985). *Competitive Advantage*. New York: Free Press.

Schunn, Christopher (2008). "Integrating Social and Cognitive Elements of Discovery and Innovation." In National Science Foundation (Ed.) (Vol. Award number 0830210).

Sturgeon, Timothy and Clair Brown (2008). "National Organization Survey by Extending the Sample and Adding an Innovation Module," presentation at the 2008 Kauffman Symposium on Entrepreneurship and Innovation Data, Washington, D.C. November 21, 2008

Sturgeon, Timothy, Frank Levy, Clair Brown, J. Bradford Jensen; and David Weil (2006). *Why We Can't Measure the Economic Effects of Services Offshoring: The Data Gaps and How to Fill Them.* MIT Industrial Performance Center Services Offshoring Working Group Final Report. Available at http://web.mit.edu/ipc/www/IPC_Offshoring_Report.pdf.

Thomas, J., and Cook, K. (2005). "Illuminating the Path: The Research and Development Agenda for Visual Analytics."

Zucker, Lynn D. and Michael R. Darby (2007). "Star Scientists, Innovation, and Regional and National Immigration," NBER working paper #13547 (October).