Chapter 4
On the Challenges of Collaborative Data Processing

Sylvie Noël
Communications Research Centre, Canada

Daniel Lemire
UQAM, Canada

ABSTRACT

The last 30 years have seen the creation of a variety of electronic collaboration tools for science and business. Some of the best-known collaboration tools support text editing (e.g., wikis). Wikipedia’s success shows that large-scale collaboration can produce highly valuable content. Meanwhile much structured data is being collected and made publicly available. We have never had access to more powerful databases and statistical packages. Is large-scale collaborative data analysis now possible? Using a quantitative analysis of Web 2.0 data visualization sites, the authors find evidence that at least moderate open collaboration occurs. The authors then explore some of the limiting factors of collaboration over data.

INTRODUCTION

Electronic collaboration tools are widespread. Many of these tools are aimed at supporting either group meetings (brainstorming tools, shared whiteboards, videoconferencing tools) or collaborative writing (wikis). These tools have been studied extensively (Pedersen et al., 1993; Okada et al., 1994; Adler et al., 2006). However, although more and more data is being collected, indexed and made available to all, collaborative data processing has received little attention until recently (Viégas et al., 2007, 2008).

Data analysis is a complex but structured task requiring specialized tools such as spreadsheets or statistical packages, some basic knowledge of statistics and information technology, and the domain knowledge to interpret the results. As opposed to text, scientific or business data is often organized in rigid structures (e.g., tables, lists, networks) and it may be more difficult to interpret without appropriate visualization tools. Regardless of these limitations, as the scientific community increasingly faces larger and larger datasets, it is time to consider techniques that may help us better understand this data.
of these difficulties, people are interested in viewing and understanding this data. Already people have access to and are familiar with financial and meteorological data, which appear regularly on television, in newspapers and on popular news sites. People are also willing to explore other types of data. For example, a website presenting statistics about baby names proved very popular (Wattenberg, 2005). Businesses of all sizes, governments, and academics analyze data for many purposes: financial planning, sales and marketing, stocks analysis, scientific research, and so on.

In companies, work-related data is called business information. The term “Business Intelligence” (BI) refers to the techniques used to improve decisions by collecting and aggregating business information. BI systems typically use a data warehouse: a large collection of historical and current data on business operations. End-user BI tools include static reports, spreadsheets linked to data repositories and interactive web applications. There is a growing business intelligence industry: the BI market grew by 10% in 2007 alone (Gartner Inc., 2007). One example of a collaborative BI business is Salesforce.com, a SaaS (software as a service) company which helps its customers share various types of business information (Dignan, 2007). Salesforce.com charges a monthly fee to customers to be able to share sales information among themselves.

While companies tend to keep their internal data private to keep an advantage over their competitors, governments and funding agencies increasingly require that scientific data repositories be accessible to all. For example, the Canadian Institutes of Health Research have a policy on Access to Research Outputs which requires grant recipients to deposit data into public databases (Canadian Institutes of Health Research, 2007). Several United Kingdom funding agencies have similar policies, including the Biotechnology and Biological Sciences Research Council, the Economic and Social Research Council, and the Engineering and Physical Sciences Research Council. In 1999, the American Congress passed circular A-110, which extended the Freedom of Information Act (FOIA) to all data produced under a funding award. China plans to make 70% of all scientific data publicly available by 2020 (Niu, 2006). There are a growing number of agencies with Open Access policies, including the U. S. National Institutes of Health, France’s Institut National de la Santé et de la Recherche Médicale, Italy’s Instituto Superiore di Sanita, Australia’s National Health and Medical Research Council, and so on. Some examples of open online scientific databases include the Generic Model Organism Database (Stein et al., 2002), the UK Data Archive for social science data, the Finnish Social Science Data Archive, and Harvard-MIT Data Center. More general open source projects for scientists are also appearing on the web. Examples include OpenWetWare.org (Butler, 2005), Science Commons (Wilbanks & Boyle, 2006), and myExperiment.org. Access to the results of scientific projects has become easier thanks to the proliferation of open access journals; the Directory of Open Access Journals (Lund University Libraries, 2003) lists over 3,000 such journals.

Open database projects also exist outside of the scientific domain, such as Swivel (Swivel Inc., 2007), Freebase (http://www.freebase.com), Numbrary (http://www.numbrary.com), and IBM Many Eyes (IBM, 2007). Amazon makes available large datasets from its web service platform (http://aws.amazon.com/publicdatasets/), including the Human Genome, various US census databases, and various labor statistics. Even the intelligence community, previously focused on secrecy, has been called to focus on information sharing (Jones, 2007). Analysis of the American intelligence efforts to prevent 9/11 has revealed that the lack of information sharing between government agencies left many of them surprised by the attack. There is a call to move from a need-to-know approach to a need-to-share one (Findley & Inge, 2005).

In spite of all this online data, we are not aware of any large-scale collaborative data analysis ini-