Comments on Proposed Guidelines for Implementing Section 515

Robert Grossman
Laboratory for Advanced Computing
National Center for Data Mining
University of Illinois at Chicago

Magnify, Inc.

August 12, 2001

Introduction

The guidelines for supporting Section 515 focus on the need to produce accurate, clear, complete, and unbiased information products. OMB has the responsibility for providing guidelines so that Federal agencies maximize the quality, objectivity, utility, and integrity of information, including statistical information, they produce.

Moreover, federal agencies are required to provide administrative mechanisms so that people affected by data which is not compliant with Section 515 quality guidelines can seek and obtain corrected information.

In this note, would like to make the following three suggestions

Whenever reports and interpreted data products are produced and disseminated by federal agencies, the cleaned data underlying these data products and reports be made available to the public.

2. That this be done by publishing the data using web based technologies.
3. That web based technologies also be used to provide the clean, but unprocessed data, as part of the administrative mechanisms provided by federal agencies to those affected by data which is not compliant with Section 515 guidelines.

Collecting, Cleaning, and Processing Data

Working with data, usually involves the following steps:

collecting data;

2. cleaning data;
3. processing data and producing data products;
4. interpreting the data products and producing reports.

Although the question of whether information is accurate, clear, complete, and unbiased is important in each of these four steps, it is particularly important in the fourth step. Traditionally, the data behind data products and reports has only been available to those analyzing it. Although reports may be widely distributed, the data behind is not.

An important step for improving the quality, utility, objectivity, and integrity of data is not only to disseminate reports, but also the underlying data behind them.

Recent advances in web based technologies provide practical, cost effective, and efficient mechanisms for doing precisely this.

The Data Web

The web today provides an infrastructure for working with distributed multimedia documents. Today, there are web based infrastructures for video (mbone), audio (napster), and distributed computing (grids). The term *data web* refers to web based infrastructures for working with remote and distributed data.

Using data webs, agencies can publish cleaned data, processed data, and data products, and the public can browse, access, and manipulate the data directly.

There are several different technologies for constructing data webs, including

- National Center for Data Mining's DataSpace
- W3C's Semantic Web
- NCSA's Data Grids
- Microsoft's .net infrastructure

There are standards and emerging standards for working with data using web based technologies, including those developed by the Data Mining Group (DMG) and the W3C.

Summary and Conclusion

Providing accurate, clear, complete and unbiased information is a challenge. Due to the difficulty of making data generally available, it has been traditional to disseminate reports about data but the not the data itself. Reports about data are more likely to be criticized for quality, utility, objectivity and integrity issues than the data itself.

Data webs refer to web based technologies for viewing, managing, and manipulating remote data. Using data webs, federal agencies can publish clean data, processed data, and data products so that the public can directly access this type of data for the first time. With the proper guidelines, publishing data on the data web can be done quickly and efficiently and is no harder than publishing documents. With common sense and reasonable guidelines, publishing data in this way will not impose unreasonable administrative burdens, and, for the first time, will empower the public with a level of information not available before.

To summarize, using the data web to publish the clean and processed data underlying the information disseminated by federal agencies helps to maximize the quality, objectivity, utility and integrity of the information.

For More Information

For more information, please contact the author at grossman@uic.edu. Information can also be found at the DataSpace and National Center for Data Mining (NCDM) web sites.

About the Author

Robert Grossman is the Director of the National Center for Data Mining (NCDM) and the Laboratory for Advanced Computing at the University of Illinois at Chicago. The Center performs research, sponsors standards, manages an international data mining testbed, and engages in outreach activities in

the areas of data mining, data intensive computing, and internet computing.

Grossman is currently the spokesperson for the Data Mining Group (DMG), an industry consortium responsible for the Predictive Model Markup Language (PMML), an XML language for data mining and predictive modeling.

Grossman is the Chairman of Magnify, a company providing outsourced data analytics, on the board of InfoBlox, a company providing network appliances, and on the scientific advisory board of the Global Information Networking Institute (GINI).

Robert Grossman became a faculty member at the University of Illinois at Chicago in 1988 and is currently Professor of Mathematics, Statistics, and Computer Science and Professor of Computer Science. From 1984-1988 he was a faculty member at the University of California at Berkeley. He received a Ph.D. from Princeton in 1985 and a B.A. from Harvard in 1980.

He has published over seventy five papers in refereed journals and proceedings on data intensive computing, data mining, high performance data management, scientific computing, and related areas, lectured extensively at scientific conferences, and organized several international conferences and workshops.