

SEMINAIRE

E+; 3=! C

SEMINAR

STATISTICAL COMMISSION AND
 ECONOMIC COMMISSION FOR EUROPE



Distr.
 GENERAL

CONFERENCE OF EUROPEAN
 STATISTICIANS

CES/SEM.43/32
 15 February 2000

ENGLISH ONLY

Seminar on integrated statistical information
 systems and related matters (ISIS 2000)

(Riga, Latvia, 29-31 May 2000)

Topic IV: Improving data dissemination strategies

**FEDSTATS: CREATING THE U.S. NATIONAL STATISTICAL INFORMATION
 INFRASTRUCTURE OF THE 21ST CENTURY**

Contributed paper

Submitted by the U.S. Bureau of Justice Statistics
 and the U.S. Bureau of the Census¹

I. INTRODUCTION

1. The Interagency Council on Statistical Policy's (ICSP) One-Stop Shopping for Federal Statistical Data web site known as FedStats (<http://www.fedstats.gov/>) opened in May 1997. FedStats has more than met its initial goal of providing the public with easy Internet access, via an initial point of entry, to the wide array of available Federal statistics. This early success is but a first step in achieving a broader vision for a National Statistical Information Infrastructure (NSII) within the United States.

2. Unlike most other countries in which there is one entity that collects, analyzes, and disseminates statistical information, the United States has developed a decentralized statistical system. For example, different agencies collect and disseminate statistical information on health, unemployment, demographics, and crime. With the rise of the Internet, however, an opportunity was presented in which a more uniform face of the United States statistical system could be presented to the public. FedStats was created to provide users with a common gateway that could quickly access the statistical information they were seeking.

3. To extend the usefulness of FedStats, new capabilities regarding integrated data dissemination have recently been implemented. With the advent of a new site feature called "MapStats," users can now more easily find statistical information related to the geographic units of states or counties found in the United States. To facilitate

¹ Prepared by Marshall DeBerry, Bureau of Justice Statistics; Valerie J. Gregg and Rachael LaPorte Taylor, Bureau of the Census.

the creation of this tool, Open Source software, such as MySQL and Perl, was used. The use of such robust and easily available tools greatly facilitated the rapid development of the requisite databases and scripts needed to provide users with the appropriate statistical information for a particular geographic area.

4. Meanwhile, FedStats serves as an incubator for detecting and/or demonstrating the challenges that lie ahead for those involved in achieving the NSII vision for the 21st Century.

II. FEDSTATS TODAY

5. The FedStats portal or gateway web site provides a centralized set of links to the web sites individual agencies have for disseminating Federal statistics. FedStats enables users to find the information they need without having to know and understand in advance how the decentralized Federal statistical system is organized or which agency or agencies may produce the data they are seeking. FedStats serves as a gateway to more than 70 Federal agencies. For more information see: *Federal Statistical Programs of the United States Government, Fiscal Year 2000*:

<http://www.whitehouse.gov/OMB/inforeg/00statprog.pdf>

6. FedStats continues to be well received and used by both the media and the public. In the first year, FedStats logged over 800,000 user sessions, and over 1,300,000 in the second year. In July 1999, Yahoo Internet Life named FedStats as one of the fifty most incredibly useful sites for the second year in a row. The fourteen ICSP agencies annually provide resources for site design, development, maintenance, and management.

7. The ICSP's Interagency One-Stop Shopping FedStats Task Force is responsible for designing, developing, enhancing and managing the web site. The task force routinely makes incremental improvements to the site by enhancing existing FedStats features. These improvements are based on user suggestions, user logs analysis, and results of on-going usability testing. The majority of the task force's work activities are undertaken by work groups, and the entire task force meets monthly to give progress briefings, share challenges, make suggestions, analyze usage statistics, etc. The task force uses a password-protected portion of the site to collaborate with each other as well as manage its work, post prototypes, meeting notes, usage statistics reports and user feedback.

8. In the process of designing and developing FedStats, the task force has encountered various approaches for disseminating Federal statistics at individual agency web sites. Part of the benefit of the FedStats initiative is that agencies learn from each other, leverage common solutions and can collaboratively seek solutions for challenges each cannot necessarily solve on their own.

9. This year's work groups are primarily focused on either developing new features or making significant enhancements to existing features. The work groups are Usability Testing, Outreach and Promotion, Data Tools, MapStats, Kid's Page, and Policy. From a site management and maintenance perspective, the FedStats Webmaster continually seeks and/or develops methods for updating the site, while technically enhancing existing features when new methods and practices become available.

10. ICSP agencies annually provide resources to cover the costs of the site's hardware, software, and Webmaster activities. The agencies also provide staff time for FedStats meetings, prototype design, development, and to conduct outreach and promotion activities.

III. CURRENT FEATURES

11. While FedStats continues to evolve, the site offers a set of core features shown below that the task force updates and/or enhances as appropriate:

- Site Map -- A handy snapshot of FedStats features.

- Search -- FedStats searches all the linked agencies (or those you specify) and then links to the documents or files it finds.
- A to Z -- Over 300 topics or subjects ranging from acute conditions (colds and influenza) through weekly earnings.
- Fast Facts -- Selected data tables from various Federal agency statistical compendia.
- Programs -- Summaries of the major statistical programs and topics of the Federal Government, including agriculture, education, energy, environment, health, income, labor, national accounts, natural resources, safety, and transportation by program and agency.
- Agencies -- A list of Federal agencies involved directly or indirectly in statistical activities.
- MapStats/Regional Statistics -- Statistics for geographic breakdowns often down to state, county and local levels.
- Contacts -- Subject matter experts for listed agencies.
- Press Releases -- For listed agencies.
- Policy -- Federal statistical policy developments including budget documents, working papers, Federal Register notices, etc.
- Additional Links -- International statistical agencies and other additional statistical resources.
- Feedback -- User comments and suggestions.

12. One of the most frequently asked questions of users is "where can I find information about my state, county, community?" The Regional Statistics Feature currently resident on FedStats is a gateway approach to finding information by geographic area on an agency-by-agency basis. However, users want to know about all the information available for a specified location, not just from one agency. That requires substantial exploration through individual agency web sites. To accommodate better those users seeking geographically based statistics, a new feature has been added, called "MapStats." Users can select a "state" or "county" from either drop-down lists or GIF images to facilitate "drilling down" to the particular geographic component. Once at the appropriate geographic level, users can then select the statistical information of interest, ranging from demographic data pertaining to population, crime, and immigration, to economic and environmental information. On selected tables, users also have the option to download the information into a format that is suitable for later analysis in either statistical software or a spreadsheet package. Future enhancements will (1) encompass additional layers of geography in which information can be made available, as well as the exploration of new technologies such as XML to facilitate the querying of heterogeneous statistical databases, and (2) integrate data sets for thematic displays.

IV. TECHNOLOGICAL FEATURES

13. The FedStats web site uses Open Source solutions in retrieving and disseminating information to site visitors. [1] Basically, open source software is software for which the original source code is freely available along with the executable program. Software can only be called "open source" if its license allows users to redistribute the program and source code at no charge. Proprietary software, on the other hand, is distributed under a very different license; it cannot be modified, copied, or redistributed without the owners' permission.

14. Loosely organized communities of programmers collaborating over the Internet

develop open source software. Advantages of using software developed in this manner include reduced cost, increased reliability, security, and support. Because of the widespread peer review, errors are found more quickly and eliminated. Code from other open source projects is reused, avoiding duplication of effort. As people have access to the source code, security holes are usually found and fixed quickly. The open source community incorporates feedback and suggestions from the users of the software, operating under the principal described by Eric Raymond, "release early, release often, and listen to your customers." [2] Technical support is provided through mailing lists and Usenet Newsgroups.

15. Some of the open source software in use by FedStats includes the Apache Group web server, MySQL relational database, Perl programming language, and Sendmail, an email program. The Apache web server is used as the site server and has operated with minimal interruptions since its installation. [3] The retrieval of information for selected data requests, such as agricultural and demographic information, uses the MySQL database for queries and storage of information. [4] The Perl language is used to parse requests for information to the MySQL database, and format the information into the appropriate HTML format. [5] Future plans include establishing a database server using Linux, an open source operating system. [6] The FedStats Task Force continues to explore new uses of Open Source software to ensure that the site will continue to operate in a reliable and efficient manner.

V. FUTURE DIRECTIONS

16. The task force is developing several new features this year. One will allow users to access in-depth information on geographic areas according to a criteria list, such as tax burden, school districts, and characteristics of the labor force. Users would then be presented with the information for different geographic locations in a manner that would allow the user to compare the differing locations. In addition, the task force continues to strive to ensure that the information presented to the public is done in a way that is easy to use and interpret. As such, the use of metadata will play an increasingly important role in the information that is disseminated to the public, so as to provide users with both contextual and attribute information about the particular statistical information they are accessing.

VI. THE NATIONAL STATISTICAL INFORMATION INFRASTRUCTURE (NSII) FOR THE 21ST CENTURY

17. The vision for the NSII is a national distributed statistical digital library with tools for information finding, for information extraction and reuse, information visualization, and for transforming knowledge into intelligence while maintaining the privacy and confidentiality of respondents. To achieve this vision, NSII will require common user interfaces, data access and searching tools usable by persons with different levels of computer and statistical literacy that enables appropriate uses of the data with analysis within and between databases.

18. The current decentralized, autonomous sources of statistical information have few commonalities in terms of concepts and definitions; system architectures, software, and hardware; measurement methods; interfaces; or dissemination and presentation modalities. Interoperability is a major hurdle in a variety of areas. Data integration issues abound. Significant challenges in high-end computing and computation and large-scale networking exist for the NSII vision to become a reality.

19. Computer and information scientists will solve some of these challenges, while others will require a more multidisciplinary, multi-sector approach. For example, involving mathematical statisticians with expertise in creating estimates from complex sample surveys, building small area estimation models, and estimating measures of error for the resulting estimates that incorporate all sources of error, including those due to sampling and non-sampling errors, should facilitate the development of multidisciplinary approaches to solving problem sets that cut across multiple areas.

20. If the metadata needed to interpret and use statistical information are to be made available and integrated with the data, the processes and procedures for collecting and compiling statistical information must also be the focus of information technologies research and development efforts.

21. The task force is confident that these issues can be addressed within the near future, and that the challenges presented in developing and providing the public with a cohesive National Statistical Information Infrastructure will be fully realized.

VII. FEDSTATS' DATA INTEGRATION RESEARCH COLLABORATIONS WITH THE NATIONAL SCIENCE FOUNDATION'S DIGITAL GOVERNMENT PROGRAM

22. Government is a major user of information technologies, a collector and maintainer of very large data sets, and a provider of critical and unique information services to individuals, states, businesses, and other customers. The goal of the U.S. National Science Foundation's Digital Government Research Program is to fund research at the intersection of the computer and information sciences research communities and the needs of government information service communities. The Internet, which was created from a successful partnership between government agencies and the information technologies research community, is a major motivating factor and context for this program.

23. There is an immediate opportunity for the broad connection of information services providers and research communities, in an arena drawing heavily on the challenging and unique requirements of the government sector, to speed innovation and development, deployment, and application of more advanced technologies into usable systems. By supporting mid- to long-term research, fundamental limitations encountered in applying information technology to the government information services domain can begin to be addressed. Research that considers real world operating constraints can provide valuable new problems and insights for the academic research domain, while demonstrating pilot systems with new capabilities for government agencies. Such research can contribute to a long-term transition strategy for migrating government information services from legacy systems, through interoperable systems of the Internet, and toward advanced integrated, global systems.

24. Within this context, the objective of the Digital Government Program is to support innovative projects that effectively and broadly address through research the potential improvement of agency, interagency, and intergovernmental operations and/or government/citizen interaction. Such research is expected to enable the generation and use of a continuous stream of advanced information technologies for early adoption and integration into the government information systems community.

25. One outgrowth of a May 1997 Digital Government workshop (see: <http://www.isi.edu/nsf/final.html>) was that the ICSP authorized a FedStats interagency Research and Development (R & D) working group. As a first step, the working group identified common challenges facing many statistical agencies that could potentially be overcome by applying cutting-edge information technologies. The working group sought academic and industrial researchers drawing from known social and mathematical scientists who are quite familiar with challenges faced by the statistical community. The working group encouraged these researchers to explore potential alliances with computer and information scientists within their institutions so that in partnership with Federal statistical agencies, they could develop and submit research proposals to several NSF research programs. In other cases, computer scientists approached the FedStats working group about possible information technology research collaborations. Agency staff spend considerable time familiarizing the researchers with current architectures, systems, and methods for collecting, processing and disseminating statistical information. They also provided explanations of the challenges that serve as barriers to achieving the NSII vision. Researchers spend considerable time

analyzing the challenges and identified possible topics that could potentially lead to interesting scientific research. Together the collaborators develop research proposals including objectives, time-lines, and any services-in-kind provided by government and/or industrial partners.

26. During the first two years of partnership building, approximately a dozen FedStats-related proposals have been submitted to various NSF research programs including the new Digital Government program (see: <http://www.interact.nsf.gov/cise/descriptions.nsf/pd/dg?OpenDocument>). This is an amazing number of successful collaborations, given the culture gaps between the partners and the amount of time all parties must contribute towards developing a competitive proposal. Most of these proposals received an award, some are still pending, and even if they do not receive NSF funding, there still has been a significant information exchange and appreciation for the expertise each brings to the partnership. Another outgrowth of these partnerships is the potential that agencies may fund portions of the research directly, in addition to or in lieu of NSF awards.

27. Over the three-year digital government award period, academic institutions collaborating with FedStats agencies will receive approximately \$4.5 million for the proposed research. In addition to the NSF award monies, several statistical agencies are supplementing the awards to ensure more of the proposed research can be completed (additional details will be provided at the Seminar).

28. The FedStats R&D working group coordinates the Federal statistical agency responsibilities and activities (along with the academic researchers) as outlined in each research proposal. The FedStats R&D working group also is fostering new and/or modified FedStats R&D partnerships that will continue to develop research proposals for submission to the wide array of NSF and other Federal agency research programs. The FedStats R&D working group is actively participating in the Federal Information Services and Applications Council (FISAC), an interagency body reporting to the Executive Office of the President's Office of Science and Technology Policy (OSTP) Technology Committee's on Computing, Information and Communications R&D Subcommittee. For more information on FISAC and other components, see: <http://www.ccic.gov/orgchart.html>.

References

- 1 See <http://www.opensource.org>
- 2 See <http://www.tuxedo.org/~esr/writings/cathedral-bazaar/cathedral-bazaar.html>
- 3 See http://apache.org/ABOUT_APACHE.html
- 4 See <http://www.mysql.org>
- 5 See <http://www.perl.org>
- 6 See <http://www.linux.org>