STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 4
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/EUROSTAT Work Session on Electronic Data Reporting
(Geneva, Switzerland, 13-15 February 2002)

Topic (i): Management, organisational and policy issues

WEB DATA COLLECTION AT THE U.S. BUREAU OF LABOR STATISTICS: AN ORGANIZATIONAL VIEW

Submitted by U.S. Bureau of Labor Statistics[1]

Invited paper

ABSTRACT

Like many other statistical organizations, the U.S. Bureau of Labor Statistics (BLS) has a significant interest in Web-based data reporting.

In 1996 BLS first launched a prototype Web system to collect establishment data for the Current Employment Survey. Though this initial prototype was made available to only a rather small number of respondents, it demonstrated the viability of Web-based electronic data reporting, and generated interest throughout the agency. As a consequence BLS decided to institutionalize Web-based data collection as an officially supported collection mode, to be made available to any program which chose to implement it.

Two key organizational decisions accompanied this choice:

- BLS chose to implement a single point of entry for all agency collection systems, regardless of program, with a uniform security model and a common look and feel across the entire site.

- BLS chose to encourage each participating program to develop its specific survey instrument on top of the shared infrastructure, recognizing the idiosyncrasies of each individual survey and avoiding a rigid, one-size-fits-all framework.

A common infrastructure and set of standards facilitates access for those respondents who supply data to multiple surveys and establishes a recognizable (branded) organizational presence. At the same time, individual surveys vary widely in scope and complexity. Differences between program areas and their approaches to data collection often have sound economic and statistical bases that go beyond simple historical accident. Thus accommodating these differences is critical.

Combining development of a centralized platform with the implementation of decentralized applications, however, has proven to be a non-trivial task. BLS has engaged in a juggling act with some effective components and some that are less so.

This paper addresses the BLS organizational approach in considerably more detail and discusses some of the trade-offs involved. In particular, the paper focuses on the roles and responsibilities required to

[1] Prepared by Michael D. Levi and Richard W. Fecher.

implement an Internet data collection facility, discusses how system development crosses organizational hierarchies, and uses the topics of security and a connection between data collection and data dissemination to illustrate some of the challenges BLS has faced.

I. INTRODUCTION

1. Like many other statistical organizations, BLS has a significant interest in electronic data reporting. The opportunities for improved data quality, reduced turn-around, and financial savings are simply too good to pass up.

2. BLS has long used several modes of electronic reporting to obtain establishment data, including telephone-based touch-tone data entry, electronic bulletin boards for file transfers, and formal Electronic Data Interchange over a virtual private network. These modes have been both efficient and cost-effective , and are now embedded in some of the biggest BLS surveys. As the World Wide Web has developed, it, too, has sparked our interest.

3. Three features, in particular, have signaled the arrival of the Web as a mature technology on which BLS can build a foundation:

- Widespread public acceptance. The pervasiveness of the Web throughout most businesses and many households means respondents are likely to have access to this medium.

- Improved interactivity. Early Web technology was relatively crude, little better than 1970's era one-screen-at-a-time mainframe CICS. Over the past several years, however, Java and other emerging technologies have provided sophisticated user interfaces with the capability for field-level and longitudinal edits.

- Improved security. Though Web cracking sometimes appears to be one of the world's most popular recreational sports, we are now confident we can establish a secure and robust architecture to protect respondent microdata and identifying information.

4. In 1996 BLS first launched a prototype Web-based system to collect establishment data for the Current Employment Survey. Though this initial prototype was made available to only a small number of respondents, it demonstrated the viability of Web-based electronic data reporting, and generated interest throughout the agency. As a consequence BLS decided to make Web-based data available to any program which chose to implement it.

5. Two key organizational decisions accompanied this choice:

- BLS chose to implement a single point of entry for all agency collection systems, regardless of program, with a uniform security model and a common look and feel across the entire site.

- BLS encouraged each participating program to develop its specific on-line questionnaire (survey instrument) on top of the shared infrastructure, recognizing the idiosyncrasies of each individual survey and avoiding a rigid, one-size-fits-all framework.

6. BLS' insistence on a shared foundation builds on our favorable experience with a unified public Web site for data dissemination. A common infrastructure and set of standards facilitates access for respondents who supply data to multiple surveys and establishes a recognizable (branded) organizational presence. BLS has now developed such a centralized capability -- the Internet Data Collection Facility (IDCF). The objective behind the IDCF project was to provide a uniform, manageable, and secure architecture for Bureau surveys to collect information over the Internet. The IDCF allows for developing, testing, and deploying survey-independent, Web-based data collection applications.

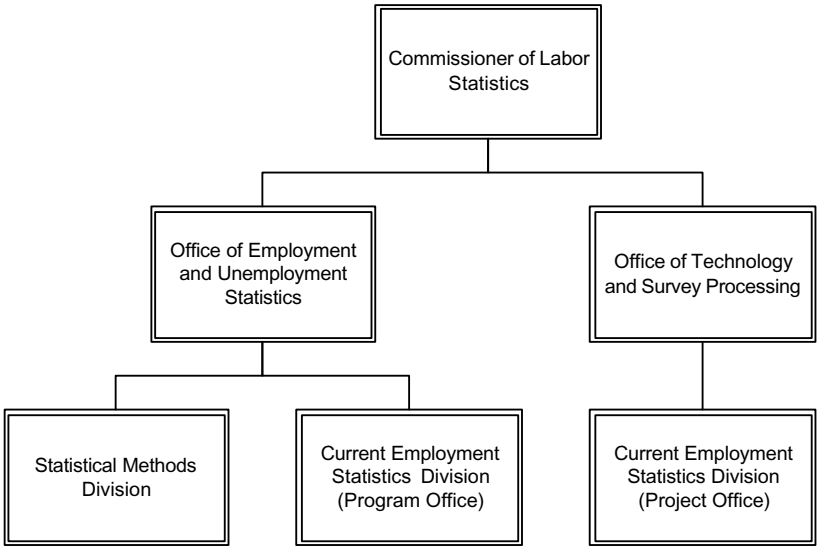
7. At the same time, individual surveys vary widely in scope and complexity, and program offices are loath to surrender control over the specifics of their collection instruments. Differences among program areas and their approaches to data collection often have sound economic and statistical bases that go beyond simple historical accident. The Current Employment Survey, for example, collects only six data items from each respondent, while the Producer Price Index collects a multi-page form with complex internal branching logic. Accommodating such differences is a critical design requirement for any collection system.

8. Combining development of a centralized infrastructure platform with the implementation of decentralized applications, however, is not a trivial task. The central challenge is one of effective communication across traditional organizational boundaries. The remainder of this paper shall address the BLS approach in considerably more detail and discuss some of the operational trade-offs we have encountered along the way.

II. ROLES AND RESPONSIBILITIES

9. The introduction of an Internet Data Collection Facility required a clear definition of roles and responsibilities among numerous groups.

10. To understand the structural impact of this decision, one needs to first understand the conventional organization of BLS survey-related software development:



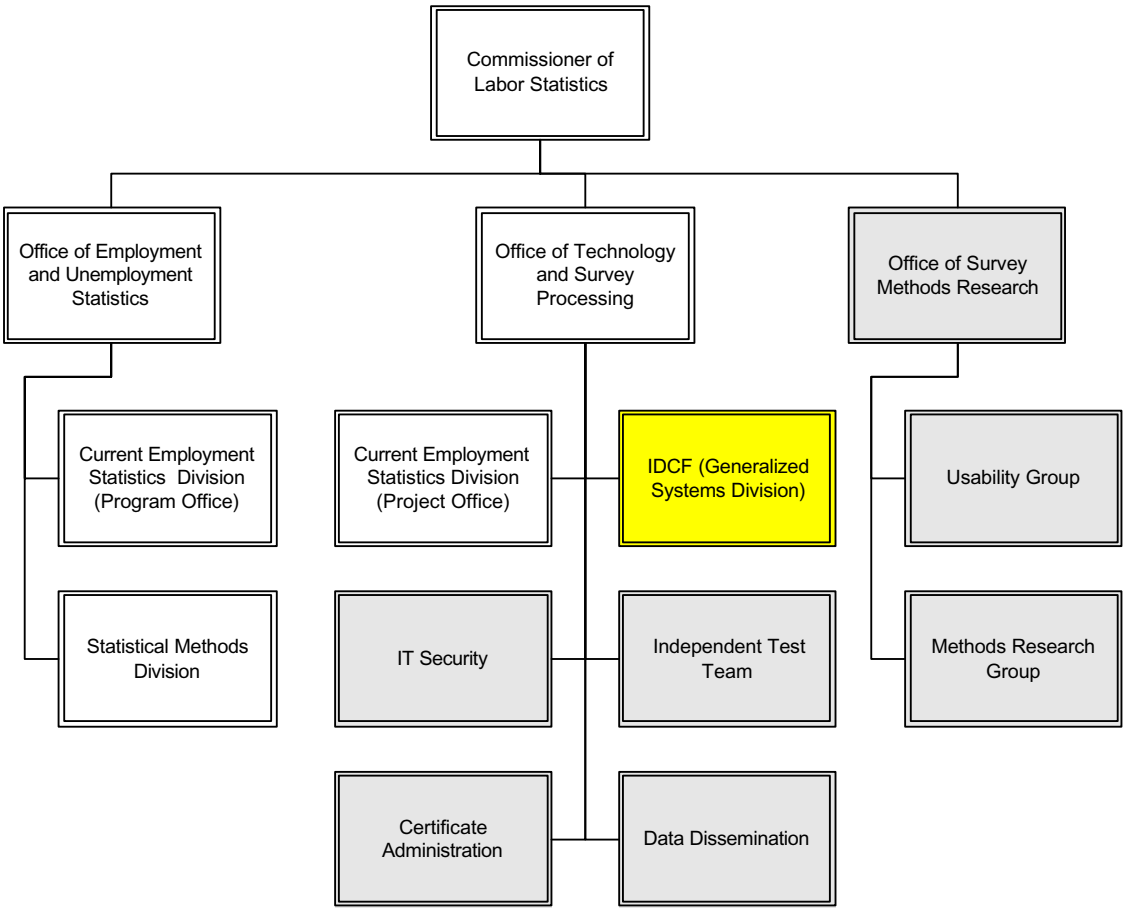
11. At the highest level, just below the Commissioner of Labor Statistics, the agency is split into Offices, each of which is headed by an Associate Commissioner who focuses on one broad program area: employment and unemployment, prices and living conditions, compensation and working conditions, or productivity and technology. These Offices are subdivided into Divisions, headed by a Program Manager, each of which is responsible for a particular survey or programs. So the Office of Employment and Unemployment Statistics has a division for the Current Employment Survey, one for the Current Population Survey, another for Local Area Unemployment Statistics, etc. These are known, somewhat confusingly, as Program Offices.

12. Statistical Methods Divisions typically report directly to the Associate Commissioner at the Office level, often with staff who specialize in a particular survey or program within that Office.

13. Computing support, both in terms of software development and maintenance as well as operations, falls under the independent Office of Technology and Survey Processing. It is divided into Divisions corresponding to the Program Offices it supports. These are known as Project Offices.

14. The fact that computing support organizations do not report directly to program managers has often resulted in internal frictions. This structure has been the institutional norm for decades, however, and staff have largely grown accustomed to it.

15. The mixed centralized/distributed nature of IDCF system development complicates this list of actors. During the project's analysis phase numerous institutional roles and responsibilities were identified, which were further refined during the subsequent lifecycle stages:



II.1 IDCF Infrastructure

Systems Manager

16. The systems manager coordinates the project. The systems manager reviews all test results and approves product implementation.

System Administrator

17. The system administrator oversees the installation, monitoring, and maintenance of all servers within the IDCF environment. The system administrator provides services such as installing, configuring, and administering hardware; granting permissions; performing routine backups; and monitoring system

and server logs. The server administrator is also responsible for the installation of shared application components such as scripts and cron jobs.

Database Administrator

18. The database administrator is responsible for allocating the database, granting and maintaining user permissions, and monitoring the database environment and engine. The database administrator is not responsible for data content or for the review of individual databases, logs, and activity.

19. While not responsible for populating the database, the administrator provides assistance to the development and production teams.

Software Development Team

20. The infrastructure development team is responsible for implementing the IDCF foundation, to include the security model, the initial log-in and validation subsystem, and all cross-survey navigation capabilities. The infrastructure development team designs and implements all application program interfaces and provides instruction and assistance in their use.

21. The infrastructure development team is also responsible for implementing and maintaining a system integration test environment.

Procedures Team

22. The Procedures Team is responsible for many activities during the application life cycle. In the development and test cycles, some of their critical activities include requirements gathering; creating test plans, test data, system documentation, reference guides, help systems, and end user support procedures; and coordinating activities among the various other teams. In the production environment their activities include supporting both end users and survey processing staff, and maintaining existing documentation.

Configuration Manager

23. The configuration manager is responsible for assuring that the project follows BLS policy regarding the management and documentation for application systems. In general, the individual is responsible for reviewing all base line documentation, and verifying that each step in the development and production life cycle of the system is reviewed and approved.

II.2 Survey Instrument

Program Manager

24. The program manager is responsible for the overall operations of the survey, including data collection activities. The program manager has ultimate authority for the published estimates.

Statistical Methods Group

25. The statistical methods group for each survey monitors collection activity from an operational viewpoint and performs the necessary analysis operations to ensure that respondent data meets survey and agency requirements in areas such as response rates, timeliness, and data integrity.

Database Owner

26. Each survey team names their database owner. The database owner is responsible for maintaining all necessary survey metadata and administering the survey microdata (loading, retrieving,

and refreshing), reviewing database logs, and monitoring activity. The survey's database owner is the primary contact for the infrastructure database administrator.

Survey Development Team

27. Each participating survey will have a core group of individuals (typically from the appropriate Project Office) responsible for developing their survey instrument. It is the responsibility of each team to design, code, and test their application components prior to the entire subsystems' placement into the survey integration test environment. Once in place, the team is responsible for performing a full integration test of their application.

Survey Librarian

28. Each survey team specifies up to two librarians who will control access and update authority for their configuration management project. Typically, the librarian is responsible for final "check-in/out" of production library members.

End-User Support Team

29. Each survey creates a support team of subject matter experts to assist respondents using the system.

II.3 Auxiliary Services

Independent Test Team

30. Composition of this team includes systems, procedures, and usability staff who are not organizationally linked to either the infrastructure group or any individual survey. The test team is responsible for assuring that the IDCF Web environment is kept current and that all proposed changes to the production system are fully regression tested. This testing and verification includes help systems and documentation.

Information Technology Security Division

31. The BLS IT security division is responsible for creating, revising, and enforcing agency security policy related to microdata confidentiality, integrity, and availability as it relates to internet data capture. The security division, in turn, reports to the BLS Security Steering Committee, composed of senior executives from both the technology and program areas.

Certificate Authority

32. One authentication method employed by BLS is digital certificates. These must be issued (and eventually renewed) through a certificate authority. Currently BLS is using an outside service for this.

Certificate Administrator

33. The Certificate Administrator is responsible for approving and revoking digital certificates and maintaining the certificate directories.

Usability Specialist

34. The usability specialist works with both the infrastructure and survey development teams. The usability specialist ensures that all user interfaces can be readily understood and effectively operated by

the intended audience, and that screens and navigation conventions are consistent across surveys and the overall production environment.

Data Dissemination Group

35. The BLS data dissemination team, though organizationally and functionally separate from data collection activities, in involved with the IDCF in two ways. First, the data dissemination group provides a small number of redirector pages on the public Web site to assist respondents in finding the IDCF and the certificate authority (long URLs often get garbled in e-mail or in telephone conversations). Second, this group is working with the IDCF Generalized Systems Division and individual Program Offices to investigate customized data displays such as comparing a specific respondent's reported data to local, national, and industry trends.

Research Group

36. Electronic data reporting research should, fundamentally, be an interdisciplinary effort. Survey methodology, computer science, information science, cognitive psychology, and cultural anthropology can be included among the disciplines which have a role in conducting such research.

37. The ultimate product of research on Web-based data reporting should be something that improves the data quality or cost-effectiveness of the organization. Quality and cost improvements can be viewed from many perspectives, however, some of which may be little-valued by any given production manager. For example, more usable software interfaces can reduce the burden on respondents who interact directly with the software, but the increase in end-user satisfaction may not be immediately visible to survey administrators. As another example, incorporating the logic or rationale used by the instrument designer into an electronic instrument could make it easier for analysts, years in the future, to understand why a specific question wording was used. Again, this is a long-term benefit which may not have an observable short-term payoff.

38. One way to enhance the breadth of knowledge being applied to a particular problem and to avoid the problems of a short-term production focus is to partner with academic researchers. This typically presents a new set of issues that need to be addressed, including academicians' lack of knowledge about the survey production environment and universities' tenure and pay decision processes.

U.S. Government-Wide Initiatives

39. All automated BLS data collection activities take place within the context of wide-ranging U.S. Federal policy and oversight. Some of the currently active areas of concern include e-government (the desire to bring Federal information and reporting vehicles on-line), accessibility to persons with disabilities (including the sight-impaired or otherwise physically disabled user community), security, and privacy.

III. IMPLICATIONS

40. The infrastructure group is a self-contained "generalized systems" organization within BLS. Though some of the units within each survey may have a different managerial hierarchy, these units are accustomed to working together. Naturally the generalized systems group and the individual survey staff must work closely together, and though this sometimes causes friction, the issues are being resolved over time.

41. One topic that has become apparent over time concerns application development tools used in different programs. Though each Program Office/Project Office pair has settled on a standard set of development tools for their survey work (programming languages, statistical analysis packages, database management systems, etc.) they rarely if ever before have needed to coordinate with other Program

Office/Project Office pairs. BLS as a whole has a strong standards program in place, but between grandfathering in legacy applications and occasionally granting exceptions where specific survey exigencies demanded, there has been a noticeable drift between program areas. These differences have surfaced as each survey interacts with the central IDCF, and are being addressed on a case by case basis.

42. Communications become particularly complex, however, when the auxiliary services staff become involved. Most of the above mentioned auxiliary services are performed by independent organizational units, so it is common for staff from all over the agency to take part in some aspect of the IDCF.

43. A case in point involved security decisions. Program staff, though of course deeply concerned with protecting respondent confidentiality, also worry about survey response rates. They believe that the overhead of Public Key Infrastructure (PKI) and the associated digital certificates may intimidate and slow down respondents, and are skeptical of any administrative burden that might discourage participation in their survey. The BLS security division, naturally, is focussed primarily on the threats of attacks on the system, be they directed towards unauthorized viewing of information, unauthorized changes to such information, or the denial of service to authorized users. They view PKI as the best technical protection against these threats. The certificate administrator, meanwhile, has become frustrated with coordinating BLS business practices with the policies and procedures of the external certificate authority. The BLS Security Steering Committee has been divided and has not issued a final decision concerning PKI use in the IDCF. The infrastructure development team is thus left waiting for a key implementation direction.

44. Another, perhaps less immediate, issue concerns linking respondents to a "keyhole" view of their impact on aggregated statistics. Some programs would like to entice respondents by providing graphical displays comparing their reported data to published estimates of local or specific industry values. The programs hope that this might be a useful mechanism to facilitate respondent enrollment and retention. Published estimates, however, are stored on a separate server which is directly accessible to the public. Sending respondent microdata to a public system for distribution over public lines, however, is a potential security risk. At the same time the IDCF group does not want to replicate services already available on the public site, nor do they wish to be distracted from their primary responsibility by developing and maintaining yet another set of tools.

45. Furthermore, staff from the BLS Office of Publications and Office of Survey Methods Research have objected to the potential impact of such functionality on data integrity. Respondents who see previous month's data for their company doing something drastically different than the industry as a whole might begin to question their reporting methods or even their business practices. In essence, the sample would become non-representative of the population as a whole in that members of the sample would possess different information, which could lead to:

- Greater asymmetries between sample members and their stakeholders (employees, customers, suppliers, shareholders) than among the broader population.
- A tendency for sample members to converge more tightly and quickly to data means.
- Estimates that look smoother than they should be.

46. Usability and cognitive testing may give some answers to these concerns, but here, too, BLS has made no final decision.

47. Perhaps the area where communication is the least developed concerns the basic research area. Systems developers and survey managers are accustomed to working on tight deadlines with clearly focussed objectives. Researchers tend to have a much broader view (though their empirical results may take narrow forms) and much looser time frames. There is general agreement that integrating research

into production activities would benefit both sides of this equation. The details of actually accomplishing this, however, often become muddled.

IV. CONCLUSION

48. Though this paper has concentrated on the disparate staff required to implement a mixed model of development, the other two candidate models have equal or greater problems.

49. Were the entire survey collection process to be fully distributed, every program office/project office pair would need to reinvent the basic infrastructure. Not only would this lead to a massive duplication of effort and resources, but past experience demonstrates that every survey would arrive at somewhat different solutions to the common challenges. Security would be tighter in some areas, looser in others. Replicating new solutions to emerging threats would be cumbersome and error-prone. And respondents who supply data to more than one survey would be forced to learn and maintain separate authorization and navigation procedures for each entity.

50. Conversely, if the entire Internet Data Collection Facility, including survey instruments, were fully centralized, there would be economies of scale but the individual programs would likely be shortchanged. What would be missing would be the body of institutional knowledge of the specific business processes of particular surveys, and the communications overhead between program office staff and the generalized systems staff regarding detailed survey requirements would likely be overwhelming.

51. Early on BLS made the strategic decision to centralize the electronic data reporting infrastructure but distribute development of the individual survey instruments. Nothing has caused us to consider reversing that decision. It makes a great deal of sense for the fundamental business processes of the agency.

52. We have, however, learned that this decision was not cost-free, and are still working to make the disparate pieces fit together.