

Distr.
GENERAL

CES/AC.71/2005/12
1 March 2005

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Bratislava, Slovakia, 18-20 April 2005)

Topic (ii): Development strategies for statistical information systems

**TOWARDS A UNIFIED PUBLICATIONS SYSTEM AT THE
U.S. BUREAU OF LABOR STATISTICS**

Invited Paper

Submitted by Bureau of Labor Statistics, USA¹

I. INTRODUCTION

1. Once upon a time the publications process at the U.S. Bureau of Labor Statistics (BLS) was relatively straightforward. Program offices would prepare their materials; these would be submitted to the Office of Publications for review; from there they would be sent out for typesetting and printing; and finally the finished publication would be distributed to interested readers – primarily through the postal service. Desktop publishing capabilities moved more of the physical production process in-house but the basic flow of artefacts remained largely intact.
2. The advent of the World Wide Web and other electronic data dissemination methods complicated this picture quite a bit. Not only were there now two different groups receiving final text (one for print publications and one for on-line distribution) but each group required different formats: camera-ready copy *vs.* HTML, respectively. What's more, customers quickly demanded more than agency-defined data presentations. Public Web users wanted to select their own data series and arrange the estimates according to their individual requirements. Thus survey data also needed to be available in an on-line database accompanied by a family of end-user query and formatting tools.
3. These changes took place over a span of years and nobody knew at the outset just where they would lead. Thus the output modules of survey production systems were tweaked to accommodate an incremental progression of enhancements. The result is an accumulation of pieced together systems, one per program.
4. This paper will describe the steps the U.S. Bureau of Labor Statistics is taking to analyze and build a unified publications system that will streamline the process described above to serve all program areas and output channels.

¹ Prepared by Michael D. Levi (levi.michael@bls.gov)

II. BACKGROUND

5. Over time, BLS has developed multiple channels to disseminate its data to end users: print (news releases, detailed reports, *The Monthly Labor Review* and other dedicated journals), the Internet (preformatted Web pages, a collection of interactive query tools that access a central data repository, flat files and Excel spreadsheets via FTP, and an e-mail subscription service), automated FAX-on-Demand, and a telephone-based interactive voice recognition system. Wireless transmission to personal digital assistants, cell phones, and other handheld devices is on the drawing board for future implementation.

6. In order to ensure a measure of uniformity to customers and to take advantage of economies of scale, these output channels are mostly administered centrally though two separate organizations. Print output still falls under the purview of the Office of Publications. Program offices typically submit text to the Office of Publications as Microsoft Word files; tables are submitted in various formats, including Microsoft Excel and TPL (Table Producing Language) output files. News releases are a special case – the Office of Publications edits each news release on paper, and the program office subsequently makes the agreed-upon changes and generates the camera-ready copy using Microsoft Word or PageMaker. Programs also submit some of their output -- including any analytical or interpretive content - intended for on-line distribution to the Office of Publications. Following review and editing of the files, the Office of Publications then transfers the files to an information technology organization, the Division of Data Dissemination Systems, which operates the BLS Web site. In other cases, programs transmit files for the Web site directly to the Division of Data Dissemination Systems. Data sets to be loaded onto the public database are delivered to the Division of Data Dissemination Systems in tab-delimited ASCII.

7. The most significant factor behind the current collection of systems is a history of independent development conducted by largely autonomous program areas. Each BLS program office has its own technology support organization which develops and operates automated production systems. Cross-program communication and cooperation tends to be informal rather than structured, reducing opportunities and incentives for identifying and pursuing joint projects. Thus each survey has developed its statistical processing systems largely in isolation.

III. PROBLEMS WITH THE CURRENT SYSTEM

8. There are two levels of duplication in the existing BLS publications flow: not only does each program have its own statistical output subsystem tied to its unique post-estimation data structures, but within each program there are parallel processes (some automated, some manual) for on-line and hard-copy dissemination.

9. The most obvious drawback of cross-program redundancy is simply the duplication of effort involved. Though initial systems development lies in the past, maintenance is ongoing. What's more, programs periodically redesign their processes and systems and thus the duplication of effort continues.

10. Another consequence of cross-program redundancy is an unnecessarily slow response to full agency implementation of new technological opportunities. Any enhancement of public access capabilities that requires even a subtle change in inputs from program offices must either be implemented incrementally as each survey in turn changes its output system, or requires a massive coordination effort to ensure that all programs complete their redesigns on a common schedule.

11. Accompanying within-program redundancy is the risk of inconsistent data reaching the public. Though rare, it is not unheard-of for a last-minute correction to be implemented in one output stream but not in the other. Thus print and on-line publications may show different data. The time lag between issuing such conflicting data and its identification and correction is an embarrassment at best and a serious disservice to our customers at worst.

12. Finally, the present system lacks unifying within-program and cross-program metadata that might be exploited to improve end-user access and comprehension.

IV. THE SEED OF CHANGE

13. Publication systems *per se* are common across most industries. The Internet is filled with publications: newspapers and journals, catalogs, brochures, annual reports, etc. What distinguishes statistical publications from the rest, however, are very large data tables with multiple nesting in both column-level headers and row-level stubs.

Table 2. Employment Cost Index for total compensation¹ for civilian and State and local government workers by industry and occupational group
(Not seasonally adjusted data)

Industry and occupational group	Indexes (June 1989=100)			Percent changes for--					
	Dec. 2003	Sep. 2004	Dec. 2004	3 months ended--			12 months ended--		
				Dec. 2003	Sep. 2004	Dec. 2004	Dec. 2003	Sep. 2004	Dec. 2004
Civilian workers	168.4	173.9	174.7	0.5	1.0	0.5	3.8	3.8	3.7
Excluding sales occupations	168.6	173.9	175.0	.5	.9	.6	3.9	3.7	3.8
Industry									
Goods producing ²	166.6	173.4	174.4	.5	.9	.6	4.0	4.6	4.7
Manufacturing	167.1	174.9	175.4	.4	1.0	.3	4.1	5.0	5.0
Service producing ³	169.1	174.0	174.7	.5	1.0	.4	3.9	3.4	3.3
Services	169.5	174.5	175.5	.6	1.3	.6	3.4	3.6	3.5
Health services	170.7	176.7	177.7	.8	1.3	.6	3.8	4.4	4.1
Hospitals	174.8	180.5	181.8	1.0	1.3	.7	4.3	4.3	4.0
Educational services	167.6	171.8	172.9	.4	1.7	.6	2.9	2.9	3.2
Public administration ⁴	168.1	174.1	175.4	.5	1.6	.7	4.0	4.1	4.3

Figure 1: Regular Statistical Table

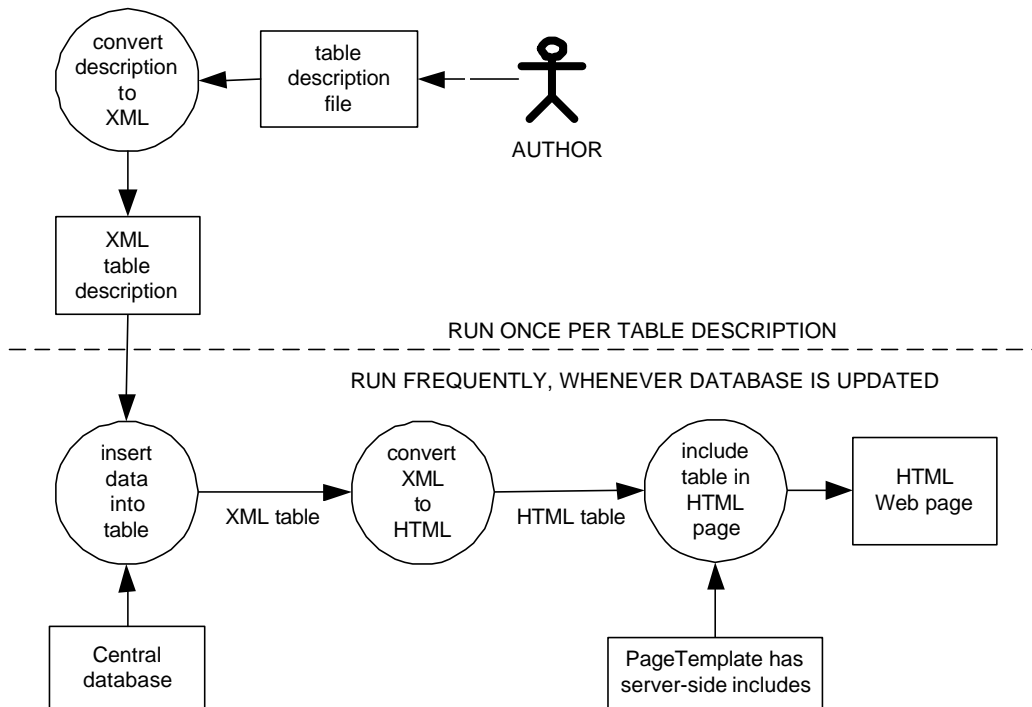
14. Commonly available commercial products rarely handle such tables well.

15. The Bureau of Labor Statistics approached this problem almost by accident, in pursuit of a different goal. Section 508 of the U.S. Rehabilitation Act requires that individuals with disabilities who seek information from U.S. Federal agencies shall have access to the information comparable to individuals without disabilities. One area in which this applies is that U.S. Federal Web sites must be comprehensible to blind or severely vision-impaired users using screen reading tools.

16. Tabular information is permitted under Section 508. The regulation states, however, that tables must be created such that assistive technologies can accurately read and render their content. This boils down to providing appropriate context information for every cell in a table.²

17. To address this issue, the Division of Data Dissemination Systems developed its Table Generation System (TGS). Using a simple table description language, subject-matter experts create table definitions. TGS then extracts estimates from the central database, merges them with the table descriptions, creates an intermediate XML file, and then converts the XML into HTML tables properly formatted to comply with Section 508.

² Stephen Ferg, "Techniques for Accessible HTML Tables" (August 23, 2002)
http://www.ferg.org/section508/accessible_tables.html



18. TGS has been running in production since 2003 on a subset of BLS tables. Its success strengthened our belief that it would be possible to develop a Unified Publication System for the entire agency.

V. VISION FOR A NEW SYSTEM

19. At the outset it should be noted that the output system described in this paper is limited to regularly occurring publications such as news releases and standard tables. Research papers, special reports, and other one-time efforts may eventually use parts of this system but they fall outside the scope of the current initiative.

20. The key elements of a unified publications system that will serve the needs of the agency as well as the customer base are:

- A central database from which all published data will be drawn;
- An XML schema that fully specifies statistical tables and publications formats (including explanatory or analytic text);
- A set of output descriptors (XML instances) that fully specify every publication issued;
- A set of transformation routines that will generate the required set of end-user products.

21. BLS already has an on-line database that contains estimates released by the agency for most current programs, going back as far as 1913. This comprises the bulk of BLS publications material. Some publications, however, go beyond estimates to include additional data such as seasonal factors, weights, and other details used to calculate the statistics. The central database will need to be expanded to hold these data, as well.

22. With the Table Generation System, BLS also has the foundation for a comprehensive XML schema. Adding features to describe text will be a fairly straightforward task. One more table type needs to be specified: what might be called a "historical" table that grows in length as new rows are appended over time

Series and year	Indexes (June 1989=100)				Percent changes for			
	Mar.	Jun.	Sep.	Dec.	3 months ended—			
					Mar.	Jun.	Sep.	Dec.
Civilian workers ¹ :								
1982	73.7	74.6	75.9	76.9	—	1.2	1.7	1.3
1983	77.8	78.8	79.7	80.7	1.2	1.3	1.1	1.3
1984	81.8	82.6	83.2	84.3	1.4	1.0	0.7	1.3
1985	85.3	86.3	87.3	88.0	1.2	1.2	1.2	0.8
1986	88.9	89.6	90.2	91.0	1.0	0.8	0.7	0.9
1987	91.9	92.5	93.3	94.1	1.0	0.7	0.9	0.9
1988	95.0	96.1	97.0	98.1	1.0	1.2	0.9	1.1
1989	99.2	100.2	101.4	102.5	1.1	1.0	1.2	1.1
1990	103.6	104.8	105.8	106.8	1.1	1.2	1.0	0.9
1991	107.9	109.0	109.8	110.6	1.0	1.0	0.7	0.7
1992	111.5	112.2	112.7	113.7	0.8	0.6	0.4	0.9

Figure 2: Historical Table

23. Finally, a few additional features need to be specified in the XML schema, the main one being the capability to suppress cells or rows when the underlying estimates are not publishable due to lack of statistical significance or non-disclosure concerns.

24. Perhaps the most time-consuming part of creating a full output system will be the creation of descriptors, or XML instances, which express in complete detail each publication to be produced. Fortunately this can be approached on a program-by-program basis, with subject-matter experts across BLS working cooperatively with the Office of Publications and the Division of Data Dissemination Systems. Past experience with TGS shows that specifying the first few tables is rather slow and error-prone, but that the learning curve picks up quite rapidly and subsequent table descriptions can be written with increasing speed and ease.

25. The biggest technical uncertainty at this point relates to the transformation routines. Generating output for on-line distribution is not a concern. TGS already creates well-formed HTML. Adding plain ASCII text as an output option will be trivial, and a separate project now underway in the Division of Data Dissemination Systems is solving the challenge of creating formatted Excel spreadsheets. Publications quality or camera-ready copy, however, may be more problematic. It is not clear that currently available commercial tools can handle statistical tables with the precision the Office of Publications requires for printed material. Under the present process a fair amount of manual adjustment is still necessary to correctly format the fine details of a printed page. Of particular concern are features such as column titles appropriately sized to correspond with data widths, and page breaks at logical row-level category boundaries with the proper stubs repeated on the following page. PDF creation tools are improving every year, however, and some recent vendor demonstrations give BLS the impression that commercial products are very close to providing the capabilities we demand.

26. These are also still some logical uncertainties regarding timing in the processing flow. Two, in particular, deserve special mention:

27. The first has to do with a conceptually vague boundary between a program's estimation subsystem and its publications subsystem. One of the methods programs use to validate estimates is to produce and review draft tables. Some of these review tables ultimately become published products. Others are for internal use only. Exactly how far into the review process a publications system should reach needs further analysis and may end up being resolved on a case-by-case basis.

28. The second uncertainty relates to control of the data. During estimation the data is clearly in the custody of the program. During output generation the data must be available to the publications staff. Exactly when the handoff occurs, and what impact this will have when data collection runs late or last-minute corrections need to be applied also requires further analysis.

29. Once the basic publications functionality has been established, numerous expansions that go far beyond current capabilities will become possible. On the print side the author expects programs to start taking advantage of easy graphing and mapping functionality and include expanded illustrations in their news releases and other publications. It is on the electronic side, however, that the most benefits will appear. Some examples include enhanced Web search capabilities; automatic generation of cross-references and suggestions for related materials; full graphing, mapping, and other data visualizations; and greatly improved cross-program end-user tools.

VI. CONCLUSION

30. Several of the essential components of a new system have been developed and successfully implemented. In particular, a central database containing the vast majority of published data is in place, a core XML schema for statistical tables has been developed and is in use, 77 table descriptors have been written and debugged, and the first set of transformations (XML plus data to HTML) has been operational for eighteen months.

31. Not all programs, however, are convinced of the need for a new system. Some of their objections include:

32. "If it ain't broke don't fix it." After all, the existing approach has successfully disseminated BLS information to the public for decades. Thus there is little urgency for change.

33. A large development effort carries with it an opportunity cost. Converting existing program output subsystems – especially creating output descriptors -- will place substantial, even if short-term, demands on program staff and their affiliated technical support organizations. This will surely interfere with other high-priority projects.

34. Finally, some programs anticipate a reduction in their control over the products that define their success and are concerned that their idiosyncratic requirements may not be fully implemented.

35. An unexpected opportunity recently opened: two major BLS programs are engaged in comprehensive redesigns of their information systems and have recognized the efficiencies an externally developed and operated publications system could provide. These two programs have jointly chartered an exploratory team to analyze and write specifications for a full replacement publications subsystem; the Office of Publications and the Division of Data Dissemination Systems are full partners in this team.

36. The author of this paper expects these two programs to provide the impetus and organizational weight necessary to bring the publications project to fruition, with full release for these programs expected no later than fiscal year 2009. Assuming the project is successful, other programs can subsequently be expected to join the fold.
