WP. 23
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Dissemination and Communication
(12-14 September 2006, Washington D.C., United States of America)

Topic (iii) How to present metadata

RESEARCH-BASED METADATA REQUIREMENTS FOR A BLS REPORTS ARCHIVE[1]

Supporting Paper

Submitted by Bureau of Labor Statistics, United States[2]

I. INTRODUCTION

1. The Bureau of Labor Statistics (BLS) is creating an Internet accessible archive consisting of publications dating back to 1886. Most of these publications are in paper form and must be scanned. All the scanned files will have appropriate metadata added to them. Currently we have only 10 years of the past 120 years of our publications available online.

2. Expectations of users are increasing for gaining digital access to historical documents. It is much cheaper to provide publications electronically via the Internet rather than adhering to older methods of mailing disks or paper copies of documents. Many of the older publications exist as a single paper copy and most are located at our national office. If the BLS national office suffers a disaster many publications will be very difficult to locate or even destroyed altogether. We want to preserve these more securely as well as make them available to the public. A solution to this situation is to digitize them.

3. A few years ago the BLS hired an outside contractor to scan some 750 publications. The scanned publications came back to us in no real order and without any metadata. It was soon decided that we needed to educate ourselves regarding the work required to produce an archive useful for both dissemination and preservation.

4. Because the National Archive and Records Administration will be responsible for the long-term preservation of our digital documents, as they are for all our permanent records, we participated in the Cornell Library's Workshop for Long-term Digital Preservation. We needed to know what is expected in long term preservation of digital objects. We also needed to determine what to consider when presenting digital objects to an open user community. After participating in a number of workshops and seminars we decided upon Adobe Acrobat's Portable Document Format/Archive (PDF/A) with Extensible Metadata Platform (XMP) as our file format and metadata technology.

5. We are near the start of the process of scanning these publications and will be using the Minolta 7000 Scanner. It has optical character recognition (OCR) capability, however we aren't planning to rely on it during this first phase of the project.. Once we have assembled a respectable collection of digital documents, we hope to disseminate them through an Internet accessible archive available through the BLS website (www.bls.gov). For this, we must develop an effective user interface, containing several ways to search for documents. Existing state government archives will be used for inspiration.

---

[1] The opinions in this paper are due to the authors and do not necessarily reflect the policies of the Bureau of Labor Statistics.
[2] Prepared by Scott C. Berridge, John J. Bosley, and Daniel W. Gillman, U.S. Bureau of Labor Statistics.

6. How we design an effective web site for the archive is a large part of the project. We have determined that we will scan all the documents but leave them as images. We wish to provide metadata for our users to help them find appropriate publications, but we are not sure what metadata is most effective. User studies, based on group interview techniques of typical users of BLS data will shed light on this problem.

7. So, in this paper, we describe in more detail some of the main issues discussed above. First we detail the scanning process and some of the other physical needs of the archive, such as disk space requirements. Next, we discuss the element set for the metadata items that are most important for the archive. In particular, the metadata needs for dissemination and preservation are compared and contrasted. Also, the importance of taking metadata items from a standard set is reviewed. Finally, the results of some user studies conducted to help determine which metadata elements are needed by users is provided.

II. ARCHIVE REQUIREMENTS

8. Disk space requirements are still not known at this time, because we don't have an estimate of the average size of the PDF files of the scanned images. We do know that there will be servers both inside and outside the firewall for staging and testing of the content of the archive and to provide access to the public. One of the luxuries we have is that all of our publications are in the public domain and do not have any copyright restrictions. Another question we have not answered is how to ensure that any publications accessed through our site are authentic. There are a number of ways to do this using the existing software, but we have not settled on one yet

9. Still, we have some time to make final decisions on these questions as it will take some time to get enough publications digitized to begin the archive. We estimate that we can scan approximately 100 pages per hour at a maximum with our resources. We do expect to have the archive ready for public access by September, 2007 when we release the updated BLS website. As a result of what we have accomplished so far we are now considering on implementing bureau wide policy regarding formats and the metadata schema and applying them to all documents produced as permanent records.

10. We are near the start of the process of scanning these publications and will be using the Minolta PS7000. The scanner’s capabilities are:

- 17" x 23-3/8" scanning area easily scans oversized bound volumes, ledgers, archival records, and other large documents
- Up to 600 dpi resolution for 11" x 17" originals, 400 dpi for 17" x 23-3/8" oversized pages with legible text and clear, sharp halftones
- Automatically compensates for page curvature at the center spine when bound volumes don't lie flat
- Automatically masks borders, erases shadows, and eliminates the images of fingers that may be holding pages open
- Scanning speed: 4.5 seconds per page (8 1/2" x 11")
- Centers scanned image for output

11. As we do not have a complete inventory, we estimate we are holding approximately 8000 publications to be scanned. Once we have assembled a respectable collection of digital objects we hope to disseminate them through an Internet accessible archive accessed from our website (www.bls.gov).

12. We want to provide a variety of ways to search and browse through the archive’s collection and are looking at developing our own architecture based upon inspiration provided by some current state

libraries. The archive is expected to be organized for search or browse for words and phrases, time period, and topic. We hope to offer a website directory, a preferences site to change the appearance of a search, and My Favorites to view publications saved in the Favorites of the archive.

13. While we have been offered a license at no charge to use the Arizona Memory Project (AMP) as the architecture of our archive, there is some hesitation to do so due to security and compatibility concerns. The BLS is very security conscious. One consideration is to create our own in-house architecture based on archives such as the AMP.

III. METADATA ELEMENTS

14. The term metadata as used here means the data used to describe a file and its contents. This is consistent with the usual informal definition of metadata, which is data about data.

15. Metadata fulfils two main needs for the archive: dissemination and preservation. By dissemination, we mean the metadata required for users to search and decide which publications are most appropriate for their needs. This corresponds to the contents of an archived file. These elements include program and subject matter metadata. By preservation, we mean the metadata needed by archivists for maintaining the archive and keeping track of files. This corresponds to the files themselves, and the elements typically contain location and size of the file, who is responsible for the file, the date the file was created, and others.

16. Searching for publications, especially on the web, is enhanced by the use of well-known tags for metadata elements. If a user knows that some standard set of metadata elements, including known tags, are used to describe the files and their contents, then this makes searching easier. An effective way to do this is to use some metadata standard. Since the publications in the archive refer to labor statistics, it makes sense to choose an element set developed within the statistical community.

17. The social science data archive community has developed a standard set of XML elements called the Data Documentation Initiative (DDI)[3]. It is used to document social science data sets, and it has much detail for describing the contents of such data. However, because of the structure of the DDI, even small subsets of the overall element set may conform[4] to the standard.

18. The BLS archive is intended to contain publications dating back to 1886. Not that much is known about the contents of these publications, and the metadata, therefore, must be relatively general in nature. In addition, the resources available at BLS to dig into the contents of each publication and thoroughly describe them is limited. Therefore, a small set of elements from the DDI was chosen to use to describe the publications. This subset conforms to the DDI, and it corresponds to the dissemination purpose of our metadata scheme.

19. In addition, descriptions of the files themselves are needed, too, and this correponds to the preservation purpose of the metadata scheme. Again, elements from the DDI were chosen. This may not be the best choice, as the DDI is not really a standard for running an archive. It was built for describing the contents of files. Until we find a more appropriate standard for this, we will continue to use the DDI. We are looking at several choices presently.

20. Our tag name of the chosen elements and a typical example from the BLS Archive are provided below. The elements are listed based on use, either for dissemination or preservation:

- DISSEMINATION ELEMENTS --
- **<Title>** Usual Weekly Earnings of Wage and Salary Workers: First Quarter 2006
- **<Subtitle>** First Quarter 2006

[3] Data Documentation Initiative (DDI) has a web site at http://www.icpsr.umich.edu/ddi.
[4] An implementation conforms to a standard if it satisfies all the requirements of that standard.

- **<ID Number>** USDL 06-696
- **<Author>** Dept of Labor / Bureau of Labor Statistics / Current Population Survey
- **<Other ID>** N/A
- **<Producer>** Office Employment and Unemployment Statistics
- **<Copyright>** Public domain
- **<Time period>** 1st Quarter 2006
- **<Collection Date>** Feb – Apr 2006
- **<Country>** USA
- **<Geographic coverage>** National
- **<Geographic unit>** US
- **<Unit of analysis>** US workers
- **<Universe>** US employed adults, 1st Quarter 2006
- **<Kind of data>** Cross-tabulation
- **<Comments>** N/A
- **<Series Name>** Quarterly Wage and Salary
- **<Series Info>** See 4th Quarter 2005
- **<Version>** N/A
- **<Holdings info>** BLS/OPUBSS, 2 Massachusetts Ave, NE, Washington, DC
- **<Keywords>** Median,
- **<Abstract>** N/A
- PRESERVATION ELEMENTS --
- **<Title>** Wage-Salary-1Q-2006.pdf
- **<File type>** pdf
- **<Producer>** Scott Berridge
- **<Production date>** 9 May 2006
- **<Production place>** OPUBSS
- **<Scan software>** Adobe Acrobat Professional 7.0
- **<Scan hardware>** Minolta 7000
- **<Depositor>** LabStat

- **<Deposit date>** 11 May 2006
- **<Distribution date>** 12 May 2006

21. The Extensible Metadata Platform (XMP) from Adobe Systems, Inc will be used to describe and store the metadata recorded for each file. The Metagrove™ system from Pound Hill Software[5] will be used to define a schema based on the elements and provide the capability of capturing the metadata for each digital image file.

IV. STUDY METHODOLOGY

A. General Description of Method

22. The research reported here collected information from a cross-section of the general public concerning the kinds of metadata elements that would be useful to them in finding statistical information in the archive and in assessing its relevance. A literature search found only scant previous work to guide this study. One of the more useful and relevant research reports that was found was based on a previous BLS-funded study by Carol Hert and Gary Marchionini (1997) [6]. Otherwise, little attention has been paid to getting inputs about metadata from non-expert users.

23. Lacking any widely-used methodological precedent, we decided to adopt (and adapt) one of the common methods for eliciting information about peoples' conceptual tools of thought: This is the focus group method, characterized by Kitzinger (1997).[7]

B. Research Participant Selection and Recruiting

24. We selected study participants from an existing BLS database of volunteers for cognitive studies of survey methodology. In order to insure that those selected would have sufficient experience with finding and using statistical information, our recruiter used a short list of screening questions during phone interviews with prospective participants to select a sub-set who were sufficiently qualified. Only those who satisfactorily passed this preliminary screening were invited to join one of the focus groups.

25. The essential characteristics that were required for inclusion in our study were:

- A positive response to whether the individual had ever looked for any statistical information produced by a US federal statistical agency.
- A judgment that the candidate's descriptions of how and why they were seeking and evaluating statistical information indicated enough familiarity with the process to qualify them for participation. This eliminated individuals whose searches were naïve and ineffective, the information sought was not really statistical, or the individual had no clear purpose in mind.

[5] Pound Hill Software - http://www.poundhill.com

[6] "Topic, geography, and time were dimensions of statistics that seemed to be important to the focus groups in terms of distinguishing between user needs. The groups indicated that many of their users wanted local data and would take higher aggregations (state, region, national) if local data were not available. The need for current as well as historical data was mentioned and there was some suggestion made that this was one way to distinguish between questions." Hert and Marchionini, *Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations*. Final Report to the Bureau of Labor Statistics, 1997, http://ils.unc.edu/~march/blsreport/mainbls.html

[7] "Focus groups are a form of group interview that capitalises on communication between research participants in order to generate data. Although group interviews are often used simply as a quick and convenient way to collect data from several people simultaneously, focus groups explicitly use group interaction as part of the method. This means that instead of the researcher asking each person to respond to a question in turn, people are encouraged to talk to one another: asking questions, exchanging anecdotes and commenting on each others' experiences and points of view.[1] The method is particularly useful for exploring people's knowledge and experiences and can be used to examine not only what people think but how they think and why they think that way."
Kitzinger, J. "Qualitative Research: Introducing focus groups," BMJ 1995;311:299-302 (29 July)

- Finally, self-directed information-seeking participants were found by asking if they had had professional assistance or help with the search task.

26. A fairly high proportion of our volunteer pool met all these selection criteria, and we were able to recruit individuals to participate in our focus groups quickly. Five such groups were conducted, each with 4 to 7 participants. This made it easier to control the discussion and maximize its relevance to the research objective.

27. The script of topics used as a guide was modified in light of accumulating experience after each group was conducted. However, the first three groups engaged in a relatively "open-ended" and highly exploratory discussion of what information attributes they would self-select as metadata elements for statistical information. The last two groups, on the other hand, were led through a much more structured discussion and evaluation of specific metadata elements that had been tentatively selected from the DDI as a good sub-set, in the researchers' judgments, to describe BLS publications.

28. All group discussions were videotaped, and the two researchers [JB and DG] who led the focus groups reviewed these tapes. They abstracted from the discussions in the first three groups the attributes and qualities of statistical information that were most frequently nominated spontaneously by our volunteers. A similar review of the tapes from the second two groups, who evaluated the selected set of DDI elements, examined the extent to which the most popular descriptive characteristics mentioned in the exploratory sessions were contained within or clearly related to some members of the DDI-based set. These two group discussions also contained sufficient material that was tied to specific metadata elements to provide deeper and more nuanced insights into our volunteers' expectations regarding the form, level of detail, and other aspects of the top-ranking elements within the set.

V. RESEARCH FINDINGS

C. Findings from Initial Exploratory Groups

29. We were confident the volunteers had some concept of metadata, however, non-expert users of statistical information, such as the participants in this study, are unlikely to understand the distinction with data. We made this assumption here in order to steer the group discussion so that the data-metadata distinction would be gradually discovered. Consequently, during the initial phase of our research, our approach to guiding the discussion of data and metadata focused on getting the participants to use their own ideas and concepts, and we tried not to "put words into their mouths." The participants narrated in their own terms the questions to which they sought answers, how they evaluated the relevance of information found to their informational needs, and their decision about the usefulness of the information (For a more detailed model of the cognitive processes here, see Bosley and Conrad, 2001[8].). In this phase, we sometimes used metadata-free examples of data as stimuli to determine what would be the best descriptive attributes for users who were seeking to understand these data. This tactic posed the conceptual distinction between data and metadata in the starkest terms. Perhaps unsurprisingly, the participants were baffled and frustrated while attempting to make sense of these examples, so that care was taken not to continue this kind of discussion for very long! Even in the discussions of metadata-free numbers, the consensus was that there were three important attributes for statistical information: topic, time, and geography. This finding agrees with Hert and Marchionini (1997).

30. When the focus of discussion moved later to real documentary reports based on BLS statistics, the members of the different groups demonstrated a common and strong interest in the same three most significant descriptors: topic, time and geography. Beyond these attributes, however, the data-metadata distinction seemed to be easily forgotten when an abundance of contextualizing metadata or narrative based on the data is supplied.

31. Some time was devoted to an exploration of the significance sample surveys. Many participants indicated an awareness of and interest in the target population, some assurance that the sample was properly drawn and representative, and other elementary methodological information. On the other hand,

[8] Bosley, J.J. and Conrad, F.G. "Usability Testing of Data Access Tools," Proceedings of the 2nd International Conference on Establishment Surveys, June 17-21, 2000, Alexandria VA: ASA Publications, pp. 971-980 (2001)

although most were also aware of sampling error and variances, they did not express much interest in specific information about error. Gaining information regarding the presence or absence of bias was sometimes identified as an issue, although only a small minority expressed this. Most seemed to trust the federal statistical agencies; those who did mention possible bias sometimes did so as part of a lack of faith in the objectivity of federal statistics. This in turn was based on an impression that some political or economic agenda was at work to influence the statistical production process.

32. Finally, an important but ancillary methodological finding is worth reporting. It is important to state that while these volunteers were paid for their participation in the study and demonstrated great willingness to "play along" with the researchers in a meaningful way, in every case the discussion revolved around describing statistical information that was of little personal, subjective interest or value to the participants. This rendered the discussions somewhat abstract or academic. As such, it was probably difficult for most volunteers to conceive of desirable metadata attributes beyond the most essential: topic, time, and geography. Nevertheless, serendipitously, there was an unexpected and important finding. Lacking *a priori* and self-generated needs for the information, many volunteers wanted information that would answer the questions: "What can I do with this information?"; or "How could this information be useful to me?"

D. Findings from Groups Focused on DDI Elements

33. The two groups whose discussions were oriented to an evaluation of the selected DDI elements strongly confirmed the primary importance of the key descriptors uncovered in first phase: topic, time, and geography. Interestingly, the title was expected to contain the topic. The value of identifying the organizational source of the information was not explicitly tested since this was a given in this case. This may have been an oversight, as a review of the earlier tapes showed that volunteers made good use of this information in the case of the metadata-free data. We cannot assume, therefore, that the informational value of source organization is negligible.

34. In their discussions of the important metadata contained in the title of the material, these groups reiterated their strong desire for some information indicating how the information can (in general) be useful. For example, one segment of discussion indicated that the inclusion of the phrase "A Guide" within the title might be nearly as valuable to users as knowing the subject-matter or topic, since this indicates that the information will incorporate some "directions for use."

35. If there is insufficient descriptive information in the title, the groups indicated that they would likely first seek additional information in the subtitle element. If that in turn was insufficient to meet their needs, keywords could be used to extend and refine topical information. The groups were, however, only willing to admit limited use of keywords. They demonstrated a fairly sophisticated appreciation of the possibility that long lists of keywords provide so much and so varied information as to become more confusing than helpful.