웹 이용자를 위한 통계 메타데이터: 통계정보 제공사이트의 메타데이터 제공 수준 평가 사례 연구

Statistical metadata for users: A case study on the level of metadata provision on statistical agency websites

오정선(Jung Sun Oh)*

ABSTRACT

As increasingly diverse kinds of information materials are available on the Internet, it becomes a challenge to define an adequate level of metadata provision for each different type of material in the context of digital libraries. This study explores issues of metadata provision for a particular type of material, statistical tables.

Statistical data always involves numbers and numeric values which should be interpreted with an understanding of underlying concepts and constructs. Because of the unique data characteristics, metadata in the statistical domain is essential not only for finding and discovering relevant data, but also for understanding and using the data found. However, in statistical metadata research, more emphasis has been put on the question of what metadata is necessary for processing the data and less on what metadata should be presented to users.

In this study, a case study was conducted to gauge the status of metadata provision for statistical tables on the Internet. The websites of two federal statistical agencies in the United States were selected and a content analysis method was used for that purpose. The result showing insufficient and inconsistent provision of metadata demonstrate the need for more discussions on statistical metadata from the ordinary web users' perspective.

초록

디지털 도서관을 통해 제공되는 정보 자원의 형태와 종류가 다양화됨에 따라 자료의 유형별로 적정 수준의 메타데이터를 정의하고 제공하는 것이 또 다른 과제로 대두되고 있다.

일반 텍스트 자료와 달리 수치로 표현된 데이터에 대한 해석을 필요로 하는 통계 자료의 특성상, 통계 도메인에서 메타데이터는 통계 자료의 검색뿐 아니라 검색된 자료의 정확한 이해와 활용을 위한 필수적인 도구로 인식되고 있다. 하지만 기존의 통계 메타데이터 연구는 통계 작성 기관이나 분석 기관의 전문적인 요구에 중점을 두고 있어, 인터넷을 통해 통계 자료에 접근하는 일반 이용자들의 관점에서의 논의는 상대적으로 부족한 실정이다.

일반 이용자를 위한 통계 메타데이터에 대한 논의의 단초로서, 본 연구는 미국의 연방 통계 기관인 the Bureau of Labor Statistics (BLS, http://www.bls.gov/) 및 the Energy Information Administration (EIA, http://www.eia.doe.gov/)의 웹사이트에 대한 내용 분석을 통해, 현재 인터넷을 통해 통계 자료에 접근하는 이용자들에게 제공되고 있는 메타데이터의 현황을 평가하였다. 본 사례 연구의 결과는 이들 웹사이트를 통해 제공되는 방대한 양의 자료에도 불구하고 메타데이터의 제공 수준은 국제 기구에 의해 정의된 최소 수준에 미치지 못함을 나타내고 있어, 이용자 중심

* School of Information and Library Science, University of North Carolina at Chapel Hill(ohjs@email.unc.edu)

의 메타데이터 설계의 필요성을 재확인 하고 있다.

Keywords: statistical data, metadata, content analysis, case study
통계정보, 메타데이터, 내용분석법, 사례연구

1 Introduction

As increasingly diverse kinds of information materials are available on the Internet, it becomes a challenge to define an adequate level of metadata provision for each different type of material in the context of digital libraries. This study explores issues of metadata provision for a particular type of material, statistical tables.

In the statistical domain, the primary representation of information always involves numbers and numeric values. A number, however, does not deliver the meaning in and of itself: it is the contextual information about the underlying concepts, structures, processes, and so on that enables people to understand what the specific number represents. In other words, statistical data should be interpreted with an understanding of the context of the data. For example, wrong assumptions about the statistical population the data at hand represent will lead to misinterpretation and false inferences. Statistical metadata in essence is about the semantic context of data, and is essential not only for finding and discovering relevant data, but also for understanding and using the data found. Due to this unique characteristic of statistical data, issues related to creating and maintaining metadata have been broadly recognized and studied, especially by parties involved in producing and disseminating official statistics such as national statistical agencies.

The increasing availability of statistical data on the Internet, however, poses a new set of challenges. First of all, the potential users on the Internet show a wider variation in their knowledge of and experience in accessing, understanding, and using statistical data than those who were typical users of statistical publication in print. In addition, the change of the medium entails changes in the way people access and work with the data. By using commercial search engines, for example, one can get a bunch of resources on ‘unemployment rate’ in the United States from across different statistical agency websites, possibly reporting quite different numbers. A problem is that statistical data that used to be physically ‘packaged’ with associated metadata (contextual information) in a printed report now can be retrieved separately. Considering these factors, it is important to make appropriate metadata visible and accessible for users in a consistent way.

In this paper, we will first review various approaches to defining and developing statistical metadata with an emphasis on data users’ perspective. The second part of this paper will report a case study conducted to gauge the status of metadata provision for statistical tables on the Internet. The websites of two federal statistical agencies in the United States were selected and a content analysis method was used for that purpose.

2 Statistical metadata

While there has been a widespread recognition of potential problems of incomplete metadata (French, 1991), there is little agreement on what constitutes an adequate level of metadata for given statistical data. The question, of course, should be answered with reference to the nature and the intended use of the data.

Broadly speaking, there are two types of statistical data: microdata and macrodata. While microdata represents observed or derived values of certain variables for certain objects, macrodata represents estimated, summary values of variables concerning sets of objects, populations (UNSC and UNECE 1999; UNSC and UNECE 1995). The basic distinction between these two types is that microdata is observed data for individual objects, whereas macrodata is aggregated or estimated data for a population or a group of individuals. In most cases, the statistical data provided by the official statistical offices, whether through the Internet or the printed medium, is macrodata in the form of the tabular report. Only limited amounts of the microdata can be accessed by outside users for secondary analyses, and even in that case the data is accessible through a closed channel rather than the Internet. Therefore, when we consider the general use of statistical data through the Internet, it is reasonable to assume that the data would be aggregated statistics (macrodata).

In most general terms, statistical metadata is about the context in which the statistical data arose. Statistical data, either microdata or macrodata, is in the form of numerical values representing the magnitude of certain attributes. The meaning of the values can be determined when they are combined with the associated metadata (Hand, 1993; Grossmann, 1999). Due to the broad range and complexity of statistical data, however, many different definitions of and approaches to statistical metadata have been proposed since Sundgren (1973) first mentioned the term "metadata in the context of statistical domain (Sundgren, 1973; cited in Froeschl, 1997).

2.1 Functional categories

The United Nations Statistical Commission & United Nations Economic Commission for Europe (UNSC & UNECE, 1995) defines statistical metadata as "data which are needed for proper production and usage of statistical data (underline added). Similarly, Grossmann (1999) suggests that metadata can be defined as all kind of information necessary for managing a statistical system and for communicating the context of the data to users and to other systems. These definitions imply two equally important users of statistical metadata: data producers and data users.

From the perspective of data producers, according to Gillman and Appel (2000), two main functions of statistical metadata are 1) to support automated survey design and processing and 2) to facilitate Internet data dissemination. However, until recently, most research has been focused on the first part, metadata for supporting automated processing. Throughout the data production process, the metadata defined/created in the previous phase often informs and guides the next phase (Sundgren 1993). In that sense, metadata is an indispensable tool for data production. Grossmann (1999) emphasizes the processing aspects by defining metadata as "information necessary for processing statistical objects in the context of statistical systems. In his view, an important goal of metadata is to support the reuse of data within a statistical system and to share the data with other systems. Klensin (1991) also takes a process-oriented view and proposes three classes of metadata for three main functions: managing data, providing information for data analysis, and managing the analysis process.

On the other hand, metadata is a supportive tool for data users. In order to be interpreted and used correctly, statistical data have to be accompanied with adequate metadata. However, in statistical metadata research, more emphasis has been put on the question of what metadata is necessary for processing the data and less on what metadata should be presented to users. As Stephenson and Clowes (1989) noted, "A problem facing users of statistical databanks is obtaining secondary information relating to the statistics. Such information includes methods of collection, problems with individual observations, precise definitions and the conceptual framework underlying the data. Drewett and the

colleagues (1989) suggest that metadata should function as an "interpretative interface to statistics and argue that this function is currently missing by stating "… the empirical researchers suffered from ignorance about the context of the data being utilized. When it was available, information about information, or meta-information, reflected the information provider's priorities. These may not have coincided with the information user's needs. Cubbitt (1990) points out that metadata is indispensable for both retrieval and analysis of statistical data: "Retrieval: how to find data, is it available, quality, appropriateness for certain types of analysis, combination with data from other sources … and "Analysis: knowing what processing can be validly undertaken on the data and what data can be compared and combined.

Recent discussions on metadata from data users' view have mostly been made in the context of Internet data dissemination. The UNSC and UNECE jointly publish a couple of guidelines and standards for statistical metadata on the Internet (UNSC & UNECE, 1998; UNSC & UNECE, 2000). Given that diverse users with different needs and different levels of statistical knowledge/experiences can access statistical data in the Internet environment, the guidelines first make a distinction between two broad groups of users: the general public who are interested in aggregate data on a general level, and the experts (researchers, subject-matter statisticians etc.) who are intensive users of detailed statistics (UNSC & UNECE, 1998). The types and amounts of metadata to be provided and the way for them to be presented should differ according to the user groups from a minimum set of metadata (e.g. measurement unit, time, geographical coverage) for correct reading of tables to extensive metadata items for detailed second analysis (the formal definition of concepts, methodologies, inclusions and exclusions in variables, relations between data objects, classification systems, etc.). As a staring point, the UNSC/UNECE guidelines specify a minimum set of metadata with respect to the usage or function of the metadata (metadata for searching data, metadata assisting interpretation, metadata for quality assessment, and metadata for post-processing).

2.2 Structural categories

According to the level of formalization and the possible system operations, statistical metadata can be classified into three categories: offline metadata, online unformatted metadata, online formatted metadata (Froeschl, 1997).

Offline metadata, also called metainformation, is all traditional documentation, usually separated from the primary data (statistical data) and presented in printed data documentation. Although this is a comprehensive reference to various aspects of data production, traditionally this kind of information is only created and managed for the data production purpose. Thus virtually no systematic access is warranted.

Online unformatted metadata, also called metatexts, is textual information accompanying the primary data. In terms of the content, it is similar to offline metadata, describing data production aspects and definitions of data objects. The main difference from the offline metadata is that metatexts allow indexing/retrieval and cross-referencing. With appropriate tools, metatext segments (documents, reference items, explanatory entries, footnotes, etc.) can be associated formally with the primary data such that an access can be made in either direction, from the data to the related metadata and vice versa.

Finally, online formatted metadata, which is called metadata in a narrow sense as opposed to metainformation or metatexts, is a structured formal description of elements closely associated with the primary data. In a strict sense from a data producer's point of view, formatted metadata is used to control data processing. Statistical metadata standards such as Data Documentation Initiative (DDI) and Statistical Data and Metadata Exchange Initiative (SDMX) fall into this category. These standards define metadata elements and structure mostly for the sophisticated needs of data archives and statistical institutions.

In another view, the above three categories of metadata can be seen as unstructured, semi-structured, and structured metadata. Similarly, OECD (2006) makes a distinction between "structural metadata and "reference metadata. While structural metadata refers to more formalized descriptions for the identification and retrieval of data, reference metadata provides methodological information including statistical concepts, information on the collection, and generation of data. It is noted that reference metadata could be generated and disseminated separately from data, and could be relevant to many instances of data at different levels, for example all data items in a dataset, or several data sets from a specific agency, etc.

3 A case study on statistical metadata provision on the Web

This section reports a case study conducted on two federal statistical agencies – the Bureau of Labor Statistics (BLS, http://www.bls.gov/) and the Energy Information Administration (EIA, http://www.eia.doe.gov/). [1]

3.1 Purpose and Scope

The purpose of this study is two fold. The first is to investigate the current practice of metadata provision with and related to statistical tables in order to assist users to properly understand and use the tables in statistical agency websites. The second is to explore the applicability of content analysis to such an investigation and to develop an instrument that can be used in a future larger study.

Broadly speaking, this study is to investigate what kinds of "metadata for "statistical data are provided to "users in the websites of statistical agencies and how they are presented. The scope of the study is determined by defining each element of the above statement.

1) Users: There is a broad range of users of statistical data, especially when it comes to the Internet environment. We define "users as ordinary internet users with limited statistical background. In other words, agency employees (internal users), statisticians, or regular users of official statistics who are well aware of the context of data production and know how to get the relevant data and metadata by their past experience are not our primary concern.

2) Statistical data: As we are mostly concerned about the use of data by ordinary users through the Internet, only official statistics (macrodata) produced by statistical agencies are of concern in this study. Among various possible formats of reporting aggregated data (table, chart, graph, etc.), we choose the statistical table because it is the most frequently used format.

3) Statistical metadata: Metadata of interest in the current study falls into the category of metatexts in Froeschl's term or reference metadata defined in OECD (2006). The rationale for studying this specific type of metadata (semi-structured, textual metadata) lies in its importance for understanding a statistical table, and especially for the users of interest in this study, ordinary users. Even though a significant body of research concerns development and application of highly structured metadata, textual metadata is indispensable for statistical data especially from the users' point of view, because this type of metadata serves well or even better to support proper understanding and use of statistical data (Sundgren & Dean, 1996). In addition, due to the high cost of generating and maintaining structured metadata, statistical data available on the Internet seldom has structured metadata.

More specifically, the following two research questions concern content aspects and

[1] The main reason for the selection of these specific agencies is in the author's familiarity with those agencies, acquired by working for the GovStat project (http://ils.unc.edu/govstat/).

presentation aspects of metadata provision respectively.

First, how much metadata is being provided in relation to statistical tables? In effect, this is about 'what' metadata is provided with a statistical table. Given the broad range of metadata, in order to assess amount or extent to which metadata is provided, we first set a set of essential metadata (see section 3.2.2 below) and examine how much metadata defined as essential is actually provided. In a sense, a comparison between 'what should be provided' and 'what is being provided' constitutes the evaluation process.

Second, how easily can a metadata element be located? Are these metadata elements presented in a way that users can easily locate them starting from the table? Is the link mechanism of the web medium used to make it easy for users to associate the metadata description with the table (and the relevant part of it)? Can a metadata instance be easily distinguished from other parts of the reference document with a label and coherence of the content?

In addition, after analyzing tables and associated metadata instances, the level of consistency across each site, both in content aspect and presentation aspect, is to be examined. In order to review the consistency in use of terms, all the link texts and labels are to be recorded along with the values of the variables.

3.2 Related studies

As mentioned in section 2, there is little research on statistical metadata from users' perspective, especially from non-expert users' perspective. Most discussions have been centered around specific issues of data producers' interest. When users of statistical data were taken into account, it was often the case that expert or professional users such as statisticians or secondary analysts were considered rather than general public users.

A notable exception is a series of metadata studies carried out in the GovStat project, a NSF-funded project for developing an integrated model of both access to and use of U.S. government statistics. The importance of providing adequate metadata in order to make federal statistics more understandable and usable for non-expert users was emphasized from the outset of the project. In order to understand what kind of difficulties users experience in using statistical agency websites, and to identify what metadata elements are necessary for users to find and understand statistical data, a series of metadata user studies were conducted (Denn et al., 2003; Hert et al., 2004; Hert et al. 2007). For the user studies, a set of tasks were developed to represent information problems that actual users would bring to statistical web sites using a scenario-based design. Three different groups of users in turn completed the tasks: data producers from within the statistical agencies, professionals outside the agencies who make regular use of the data, and undergraduates. From this mix of users it was attempted to gain an understanding of the differences between the data producer and data user perspectives, as well as the differences between expert users and novice users. Qualitative analyses of transcripts of the study sessions (in which subjects completed the given tasks) and the interviews showed that users of statistical agency websites did indeed encounter many problems due to the lack of adequate level of metadata. Among others, time and geography turned out to be the core of many user problems. For example, users had difficulties finding out what dates or periods a particular table they were looking at was reporting, or which table presented the most current statistics among many similar tables retrieved. Another important finding was that the extent of difficulties users experienced was largely related to the amount of prior knowledge they had about particular agencies, their surveys and censuses, etc. Experienced users could address problems more easily either based on their prior understanding of the context of the data or by knowing where to look up and find related metadata. Novice users often found it difficult to figure out which paths (links) should be followed to get to the appropriate metadata. More importantly, in some cases they did not

realize the need of metadata for the proper interpretation/comparison of the data at hand. This finding suggests the importance of making the association between statistical data and metadata clearly visible for users.

Although the above studies indicate that the uncertainties users experienced could be potentially resolved with appropriate metadata and that it was difficult for users with no prior knowledge to find the metadata, it is not clear whether it was because the metadata did not exist in the website or because users could not locate it. This study attempts to answer the question by investigating the level of metadata provision in both the content and the presentation aspects.

3.3 Methodology

The investigation methodology is a content analysis (Krippendorff, 1980). The analysis was carried out by the author without another coder, and therefore inter-coder reliability could not be measured. Although it weakens the study, given the lack of existing instruments for this kind of study in the statistical domain, this study was decided to be exploratory. In fact, one of the purposes of this study is to test the applicability of the method and to gain insight for further development of a research protocol.

3.3.1 Unit of analysis

- The primary unit of data collection and analysis is defined as a single html page containing a statistical table. This unit is called 'table page'.
- All documents of any text format (html, pdf, doc, etc.) within two links from a table page also form another unit of analysis.

There are several file formats being used for statistical table, including plain text (TXT), Spreadsheet (XLS), and PDF. However, as our research question concerns how much metadata is provided with tables leveraging link mechanism of the web, we only choose the html format tables for the analysis.

This constraint on file format does not apply to linked files, and all linked documents are to be examined regardless of their file formats for locating metadata instances. Two distinct coding units are defined for a metadata instance: 'reference document' and 'metadata unit'. 'Reference document' is defined as the document in which the metadata instance appears. 'Metadata unit' is defined as the smallest labeled text segment that contains the metadata instance. A metadata unit can be a paragraph starting with a label, a text block with a few paragraphs preceded by a heading, a section with a section heading, or a whole html page if the page consists of free-flowing text without any structural format (in this case, reference document = metadata unit). If the metadata item is in the same page with the table (i.e., in the footnote), the reference document and the table page is the same.

In this study, metadata provision is defined in terms of access methods for users to get to relevant metadata from the table at hand. Access methods are classified into three broad cases. The first case is that metadata is presented on the same page as the table. The second case involves hyperlink mechanism. A metadata element is considered to be provided if it can be located by following any one of the links from the table page. The third method is 'see reference.' In this case, the metadata description is not on the same page with the table, and no hyperlink path from the table page to the reference page is provided.

3.3.2 Metadata category and codebook development

The UNSC and UNECE published a guideline for statistical metadata on the Internet (UNSC and UNECE, 2000) [2], which aims to establish a "minimum standard set of metadata

[2] UNSC and UNECE published the first version of the guideline in 1998 (UNSC and

elements that should be presented with tables on the Internet. As this guideline fits well to the scope and purpose of this study, it was decided to use this guideline to define the metadata category for this study. The minimum set of elements defined in UNSC and UNECE (2000) is shown below.[3]

Minimum set of metadata required for the correct interpretation of statistics:
- Title/content description (depending on subject area/content);
  Often including the following elements:
  Statistical population;
  Geographical coverage;
  Observation unit;
  Classifications and standards applied;
- Labels for rows/columns in the tables and elements of the graphs;
- Definitions of labels;
- Measurement unit;
- Time reference/period;
- Regional units;
- Comparability over time (break in series, missing data);
- Footnotes highlighting specific precautions;
- Source of the data (agency compiling the data);
- Explanation of standard symbols in tables;
- Any information on copyright, restrictions of usage;
- Contact points for additional information;

Having metadata categories in place, the next step was to define variables and measures covering both content and presentation/organization aspects.

Prior to the sampling of each site, a coding scheme (including a codebook and two coding forms) was developed and tested with two BLS tables and one EIA table. Only minor changes were made during this process. The codebook includes basic instructions, definitions of metadata categories, definitions of variables defined in this study, and coding rules (see Appendix). The UNSC and UNECE guidelines which specify the minimum set of metadata elements being used in this study do not provide definitions of each element in the set. In order to prepare the definitions included in the codebook, two prominent statistical metadata standards, Data Documentation Initiative (DDI) and Statistical Data and Metadata Exchange

---

UNECE, 1998) and after getting comments from international organizations, national statistical agencies, etc. through UNSC and UNECE Work Session on Statistical Metadata in the Conference of European Statistics, it published a revised version in 2000 (UNSC and UNECE, 2000).

[3] UNSC and UNECE (2000) defines three different types of metadata and suggest a minimum set of elements for each type. Those three types are “metadata assisting search and navigation , “metadata assisting interpretation , and “metadata assisting post-processing. The box shows only “metadata assisting interpretation.

Items included in “metadata assisting search and navigation are not about specific tables but about the services or contents related to the whole site. For example it includes “site map/table of contents of the web site , “site news , “description of statistical institution , etc.

As for “metadata assisting post-processing , they do not specify metadata elements. The guideline says “metadata listed under the 'Minimum set of metadata required for the correct interpretation of statistics' … should either be attached to the downloadable data, or should be easy to access and download in a separate operation.

Initiative (SDMX), were consulted.

3.3.3 Sampling & coding

Because of the dynamic nature of the web environment, we decided to crawl each agency's website and store them (preserving their original link structures) in a local device. All analysis was done on the stored local copies. The copies used in the analysis were crawled on April 11, 2006.

As one of our research questions concerns the consistency in provision of metadata throughout an agency website, it was important to select samples that represent the breadth of the whole site as much as possible given the limited number of samples. To accomplish that, a preliminary analysis on the content and structure of each agency site was done.

The BLS site structure is essentially based on the survey programs and the survey program home pages serve as major nodes. Further examination of the BLS directory structure revealed that even though each program has its own directory (ex. http://www.bls.gov/iif/ for Injuries, Illnesses, and Fatalities (IIF) Program) and most files related to that program are physically located under that directory (http://www.bls.gov/news.release), news release files (which contains tables in the html format) from all the different programs are collected under one physical directory, regardless of the program to which it belongs. Thus, drawing table samples for the BLS site got down to selecting files from a single directory. A numbered list of all the files (total 525) in the news-release directory (www.bls.gov/news.release) was prepared. From the list of 525 files, 20 files were randomly selected. A website providing a random number generator (http://www.random.org) was used to generate random numbers for this purpose. The 20 files were then examined and if two or more tables were from the same program, only one table was been selected. In all, 14 table pages were selected.

EIA samples were selected using a different process. Instead of figuring out where to draw samples by exploring the site, we started from a list of html files supposedly containing a statistical table. A shell script to detect a statistical table within a html file based on heuristic rules was written and used to prepare a numbered list. From the list of 3,502 files, 20 files were randomly selected using a random number generator. Finally, the list of EIA surveys (http://www.eia.doe.gov/oss/forms.html) was used to check whether the samples evenly cover the range of data provided in EIA. As a result, 12 table pages were remained.

For each table, all the documents, regardless of the file format, within the two links from the table were examined for the presence of metadata instances. In addition, if there was a see-reference without a hyperlink, the referred document was retrieved and included for the analysis. In the case of BLS tables, since each and every page containing a table had a link to the homepage of the relevant survey program, all the documents directly linked under the survey program homepage were in effect included for the content analysis. The total number of documents examined from the BLS site was 291 (112 pdf files, 179 html files). In the EIA site where the overall design was less consistent, some tables did not have any link at all and thus the total number of documents (other than the actual table pages) examined was much smaller (53 in total).

Using the coding scheme developed for the study, all the analysis and coding was done by the author. Since the coding was done without cross-checking by other coders, reliability of the coding could not be evaluated.

3.4 Results

3.4.1 The content aspect

The number of metadata instances and types for BLS tables and EIA tables is shown in Table

1 and Table 2 respectively. On average, a BLS table had 9.14 metadata instances of 6.21 distinct types. Another way of looking at this is that among 14 minimum metadata element types, about six types of elements are provided with a BLS table, with each having about 1.5 instances. EIA tables have less; 6.58 metadata instances of 5 types were found on average.

Table 1. Number of metadata instances and types for BLS tables

Table ID	Instances	Types
BLS001	7	7
BLS002	9	7
BLS003	6	6
BLS004	5	4
BLS005	10	5
BLS006	5	5
BLS007	5	5
BLS008	13	7
BLS009	11	8
BLS010	7	6
BLS011	18	10
BLS012	11	6
BLS013	10	4
BLS014	11	7
Total	128	87

Table 2. Number of metadata instances and types for EIA tables

Table ID	Instances	Types
EIA001	12	5
EIA002	4	4
EIA003	3	3
EIA004	4	1
EIA005	11	6
EIA006	5	5
EIA007	10	8
EIA008	5	5
EIA009	8	7
EIA010	6	6
EIA011	7	6
EIA012	4	4
Total	79	60

Table 3 shows the number of metadata instances for each metadata category and the number of tables having one or more instances of that category (type). Both BLS data and EIA data show that most tables have one or more metadata instance for ‘MC05 Definitions of column/row labels, variables, etc.’, ‘MC 11 Source of data’, and ‘MC14 Contact point’. For one metadata category, ‘MC13 Information on copyright’, no instance was found in either BLS or EIA. Considering that all those tables are provided in the website of a single agency while the minimum metadata set is defined by an international institution which gathers data from across different national statistical agencies, absence of this type of metadata is understandable. The relative lack of ‘MC02 Geographical coverage’ and ‘MC08 Regional units’ metadata instances could be problematic, given that metadata user studies for the statistical domain (Hert & Hernández, 2001; Denn et al., 2003) showed the importance of geographic information.

Table 3. Number of metadata instances and types for metadata categories

Metadata Category	BLS		EIA	
	Instances	Types	Instances	Types
MC01. Statistical population	12	10	2	2
MC02. Geographical coverage	4	3	2	2
MC03. Observation unit	2	1	0	0
MC04. Classification and standards applied	9	7	1	1
MC05. Definition of colum/row labels, variables, etc.	35	11	13	6
MC06. Measurement unit	6	4	20	9
MC07. Time reference/period	9	8	4	4
MC08. Regional units	1	1	0	0
MC09. Comparability over time	5	4	3	2
MC10. Specific precautions	11	7	8	8
MC11. Source of the data	13	12	11	11
MC12. Explanation of standard symbols	5	5	8	8
MC13. Information on copyright, restrictions of usage	0	0	0	0
MC14. Contact points for additional information	16	14	7	7
Total	128	87	79	60

3.4.2 The presentation aspect

Table 4 shows the proximity measures of the 128 BLS metadata instances and 79 EIA metadata instances. In EIA, most of metadata instances were found on the same page with the table, in the form of a footnote. The 11 instances found by following the link are, four instances of 'MC14 Contact points', two instances of 'MC05 Definition of column/row labels, variables, etc', 'MC11 Source of data', and one instance of 'MC 10 Specific precautions'. In the case of BLS, most metadata instances were found by following links. Consistently throughout the BLS site, tables and documents related to a survey program are placed under the homepage of the survey program, and every page has a navigation bar in the upper right part of the page. The shortest path from a table to related metadata such as a description of the population of the survey often involves a link from the table to the survey homepage and then a link from the survey homepage to a reference document.

Table 4. Proximity to the table

Distance	BLS	EIA
The same page: A metadata instance found in the same page with a table.	45	68
A direct link: A metadata instance was one click away.	5	9
Two links: A metadata instance was two clicks away.	78	2
Other	0	0
	128	79

However, most hyperlinks are provided at the document level and in order to locate relevant piece of information within the document, it is often necessary to skim through the whole document. Table 5 shows the granularity of links.

Table 5. Granularity of links

		BLS	EIA
Direct Link	From a part of the table to a metadata unit (i.e. from a specific part being described (e.g. column heading) to a relevant text segment within a reference document; an anchored text link)	0	0

	From a part of the table to a reference page (without an anchor to the relevant metadata unit)	0	1
	From a table page to a metadata unit (i.e. from somewhere in a table page outside the actual table (e.g. a navigation bar on top of the page) to a relevant text segment in a reference page)	0	0
	From a table page to a reference page	5	8
Subtotal		5	9
Two Links	Outlink from a part of the table & Inlink to a metadata unit	0	0
	Outlink from a part of the table & Inlink to a reference page	0	0
	Outlink from a table page & Inlink to a metadata unit	10	0
	Out link from a table page & Inlink to a reference page	68	1
Subtotal		78	1
Total		83	10

Modularity of metadata instances was reasonably high in both BLS and EIA. Most metadata units consist of one coherent paragraph, not requiring users to read through big chunks of non-relevant text.

However, in order to allow users to quickly locate the relevant metadata unit, it is desirable to provide consistent labels for each metadata element across the site. Most of the EIA metadata instances are in footnotes of the tables. It is rare for a footnote to have a label. In BLS, labels for a specific metadata element are reasonably consistent with few exceptions. For example, among 11 instances of 'MC01 Statistical population', 6 instances were found under the label 'Coverage.' Some other labels for these categories do not use the term 'Coverage' but labels deliver similar meanings. For example, population for table ID BLS012 was found under 'What kind of information does the MLS program provide?' Another example of consistent usage of the label is found in 'MC05 Definitions'. Among 35 instances of this category, excluding 19 cases where metadata instances were found in the footnote of the table, almost every instance was found under the label 'Definition' or 'Definition – Term' (e.g. Definitions - Wage and salary workers).

3.5 Discussion

In this study, we examined two federal statistical agencies' websites in order to assess how much metadata is available for users to easily locate when they are using a table. It was found that among the set of 14 'minimum' metadata recommended by an authoritative international body, only about 6 elements are available for a table in BLS and 5 in EIA on average. In both sites, information on the source of the table and contact information for further inquiry is consistently provided. On the other hand, geographical information including geographic coverage and units is not provided in most cases. Definitions for statistical population, variable, and concepts are provided only on a limited basis.

Although the analysis was focused on the presence or absence of metadata elements, it was noted that in many cases the descriptions found were too succinct. Especially in the EIA site, most available instances of metadata under these categories are found in the footnote and in most cases what is provided is information on inclusion or exclusion of specific groups rather than a full description or explanation. It is hardly considered as

sufficient for ordinary internet users with limited knowledge and experience in statistics in general and statistics available on that site in specific.

In presenting metadata, it would be ideal if each metadata element was described in a distinct coherent text segment with identifying labels, and was linked to all relevant tables at the right level of granularity (e.g. a variable description to a column heading). Effective use of the hyperlink mechanism of the web combined with the composition of textual metadata in a consistent style would make that possible relatively easily. Even though the granularity of links being provided in both sites turned out to be coarse (i.e., page level),, text segments containing metadata instances seem to be composed of small coherent text with relatively consistent labels, especially in the BLS site. By virtue of the consistent design across the whole site, it is relatively clear to see what is available in the BLS site.

As this preliminary research uses only a small number of samples (14 from BLS, 12 from EIA), no generalizations to the rest of these agencies' sites, or to other agencies, can be made. In addition, there are other limitations that might bias the results. First, as we only draw samples in one format, html file, the samples are not representative for all the tables provided in those sites. In the BLS site, in almost every survey program homepage, a number of tables are under the 'Tables Created By BLS' subsection. These tables are excluded because of its format (txt file) and the location (ftp server). In case of EIA, the restriction on the file format and the location might affect the result more, because some of the EIA tables are only published in PDF file, XLS file, or both. For example, Monthly Energy Review which is one of the most important publications of EIA with large volumes only available in PDF or XLS.

4 Conclusion

This study was conducted with two main purposes in mind. The first was to evaluate the current level of metadata provision in both content and presentation aspects, and the second was to develop an instrument for a content analysis that can be applied to such an evaluation.

The result of the content analysis of two federal statistical agencies' websites showed that among the set of 'minimum' metadata elements recommended by an authoritative international body (UNSC and UNECE), only less than half of the elements were available within two links in both websites. In terms of the presentation, it was found that both websites were not taking advantage of the hyperlink mechanism. Metadata instances seldom linked to the relevant part of the table, leaving the task of relating data and metadata to the users.

The validity of content analysis depends largely on creating appropriate categories. The codebook developed in this study was based on a guideline published by the UNSC and UNECE. Since the guideline was the result of extensive discussions, within the statistical metadata community, on what constitutes a minimal set of metadata in the context of the Internet, we believe it was appropriate to adopt the guideline. Overall, a content analysis method appears to be an adequate tool to assess the level of metadata provision on a statistical website. While the current coding scheme was developed to evaluate the presence or absence of necessary elements of statistical metadata, it can be extended to measure the level of completeness of the information in a future study.

Bibliography

Cubbitt, R. 1990. "Metadata in Statistics: The Problem for Producers. In *Proceedings of Workshop on Expert Systems and Artificial Intelligence: The Need for Information about Data*, ed. M.C. Fessey, 7-13. London: The Library Association.

Data Documentation Initiative. *About the specification.* http://www.icpsr.umich.edu/DDI/codebook/index.html (accessed April 27, 2006).

Data Documentation Initiative. *Technical information*. http://www.icpsr.umich.edu/DDI/dtd/index.html (accessed April 27, 2006).

Denn, S.O., S.W Haas, and C.A. Hert. 2003. "Statistical metadata needs during integration tasks. In *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop*, 81-90. http://www.siderean.com/dc2003/301_Paper50.pdf (accessed March 23, 2006).

Drewett, R., E. Tanenbaum, and M. Taylor. 1989. "Creating a Standardized and Integrated Knowledge-Based Expert System for the Documentation of Statistical Series. In *Proceedings of Seminar on Development of Statistical Expert Systems in Computational Statistics*, ed. EUROSTAT, 178-189. Luxembourg: Office for Official Publications.

French, J. C. 1991. "Support for Scientific Database Management. In *Statistical and Scientific Databases*, ed. Z. Michalewicz, 51-82. Chichester: Ellis Horwood.

Froeschl, K. 1997. *Metadata management in statistical information processing: a unified framework for metadata-based processing of statistical data aggregates*. New York: Springer.

Gillman, D., and M. Appel. 2000. "Statistical metadata research at the Census Bureau. In *1999 Federal Committee on Statistical Methodology Research Conference*. [cited 2006.11.26]. <http://www.fcsm.gov/99papers/gillman.pdf>.

Grossmann, W. 1999. "Statistical Metadata Tutorial. [cited 2006.3.2]. <http://www.univie.ac.at/dac/papers/Tutorial_gesamt.pdf>.

Hand, D. J. 1993. "Data, metadata and information. *Statistical Journal of the United Nations Economic Commission for Europe* 10(2): 143-152.

Hert, C.A. and Hernández, N. 2001. "User Uncertainties With Tabular Statistical Data: Identification And Resolution: Final Report to the United States Bureau of Labor Statistics. [cited 2006.3.2]. <http://ils.unc.edu/govstat/fedstats/uncertaintiespaper.htm>.

Hert, C.A., S. O. Denn, and S. W. Haas. 2004. "The Role of Metadata in the Statistical Knowledge Network: An Emerging Research Agenda. *Social Science Computing Reviews* 22(1): 92-100.

Hert, C. A., S. O. Denn, D. W. Gillman, J. S. Oh, M. C. Pattuelli, and N. Hernández. 2007. "Investigating and modeling metadata use to support information architecture development in the statistical knowledge network. *Journal of the American Society for Information Science and Technology* 58(9)**:** 1267-1284.

Klensin, J. C. 1991. "Data Analysis Requirements and Statistical Databases." In *Statistical and Scientific Database*, ed. Z. Michalewicz, 35-49. New York: Ellis Horwood.

Krippendorff, K. 1980. *Content analysis: An introduction to its methodology.* Newbury Park, CA: Sage Publications.

Neuendorf, Kimberly A. 2002. *The content analysis guidebook.* Thousand Oaks, CA: Sage

Publications.
OECD. 2006. *Data and Metadata Presentation and Reporting Handbook*. [cited 2007.6.1]. <http://www.oecd.org/dataoecd/46/17/37671574.pdf>.

Petrakos, M., Farmakis, G., Koumanakos, G., and Bisiela, P. 2001. "A structured approach to a statistical metadata base. *NTTS 2001 International Conference in Official Statistics*, Creta 2001.

Statistical Data and Metadata eXchange. *SDMX Standards Version 2.0*. [cited 2007.3.1]. <http://sdmx.org/index.php?page_id=16#package>.

Stephenson, G., and I. Clowes. 1989. "Knowledge Interrogation System for Social and Economic Statistics. In *Proceedings of Seminar on Development of Statistical Expert Systems in Computational Statistics*, ed. EUROSTAT, 210-220. Luxembourg: Office for Official Publications.

Sundgren, B. 1993. "Statistical Metainformation Systems – Pragmatics, Semantics, Syntactics. *Statistical Journal of the United Nations Economic Commission for Europe* 10(20): 121-142.

Sundgren, B., and Dean, P. 1996. "Metadata: A Quality Element in Official Statistics - the Swedish Approach. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 45 - 54.

UNSC and UNECE. 1995. *Guidelines for the modeling of statistical data and metadata*. [cited 2006.2.10]. <http://forum.europa.eu.int/irc/DownLoad/kgeqAKJ2mRG0dYT2R-SC2wUfSt04mSytde1Bq7yq6HUT-SapeSGGkeZTyMxFTE_U/guidelines.doc >.

UNSC and UNECE. 1998. *Guidelines for Statistical Metadata on the Internet*. [cited 2006.2.10]. <http://www.unece.org/stats/documents/1998/02/metis/23.e.pdf>.

UNSC and UNECE. 1999. *Information Systems Architecture for National and International Statistical Offices: Guidelines and Recommendations*. [cited 2006.2.10]. <http://forum.europa.eu.int/irc/DownLoad/kjebAUJEm_Gtp9pEsrpFEcAxoWPHbTV-YFJRu9tYP21HNE7tEoPVPI0CpAjd9Bu-qV2O2/CESmetadataInternet.pdf>.

UNSC and UNECE. 2000. *Guidelines for Statistical Metadata on the Internet.* [cited 2006.3.22]. <http://www.unece.org/stats/publications/metadata.pdf>.

Appendix. Codebook

Unit of Analysis and Coding:

- The html pages containing sample tables.
- Documents (html pages & other formats) within two links from sample tables.

For each sample table, the instance(s) of a set of metadata elements (defined below) within two links from the table (exception: when there is a see reference from the table to a specific page, the referred page should be included regardless of the distance from the table) will be identified and recorded.

Two distinct coding units are defined associated with a metadata instance: 'reference page' and 'metadata unit'. 'Reference page' is defined as the page on which the metadata instance appears. 'Metadata unit' is defined as the smallest labeled text segment that contains the metadata instance. A metadata unit can be a paragraph starting with a label, a text block with a few paragraphs preceded by a heading, a section with a section heading, or a whole page if the page consists of free-flowing text without any structural format (in this case, reference page = metadata unit). If the metadata item is on the same page with the table (i.e., in the footnote), the reference page and the table page are the same.

Basic Instructions:

1. Retrieve the table page (the html page that has one or more statistical table reporting official statistics) from the specified URL in the part I of the coding form (Table ID is given in this form).
2. Read the table page completely.
3. Identify metadata elements (defined in Part I (No.5) of this codebook) presented in the table page, assign IDs to each instance of the elements, and record them in part II of the coding form.
4. Follow a link from the table page and read through the linked page.
5. Identify metadata elements presented in the linked page, assign IDs, and record them. Note that you need to record "metadata unit as well as "reference page
6. If there are links from the page (retrieved in 4), follow them (if it is apparent that a link would not contain any metadata for the table currently being coded, for example a link to another table, skip it). Examine the pages for the existence of metadata, and record identified instances (in the part II).
7. Repeat 4 to 6 for each and every link from the table page (except a link to the agency home page or links to outside the local copy of the domain).
8. If there is any "see reference in the table page, retrieve the referred document regardless of its file format (HTML, PDF, or else) and its distance from the table. If a link to the referred document or information on the location of the document is not provided in the table page, use search feature in the agency webpage (it means we moves to the actual agency website away from the local copy of the agency site. After the search is done, the availability of the retrieved document from the local copy should be checked out. If it is available from the local copy, the analysis should be done with the local copy), or navigate the site to locate the document.
9. Count all the instances for each metadata element and record the number in part I of the coding form.

Part I - To be filled out for each sample table

1. Coder ID

2. Table ID

Table ID starts with a agency code (BLS/EIA). Following the agency code, give each table a unique 2-digit number, beginning with 01 and proceeding upward without duplication across all tables within the agency.

3. Title of the table

4. URL of the table

Provide the full URL

If there is more than one table within the page (given the URL),

- If there is an anchor (allowing direct access to the table), include the anchor in the URL
- If not, provide a note including the number of tables on the page and specify which table (ex. 2nd table in the page) it is.

5. Number of metadata instances

For each metadata category, record the number of the located metadata instances. If no metadata instance is located for the category, fill in 0.

❖ Metadata categories and definitions

ID	Metadata Category
MC01	Statistical population · Population is the total membership or population or "universe" of a defined class of people, objects or events. There are two types of population, viz., target population and survey population. A target population is the population outlined in the survey objects about which information is to be sought and a survey population is the population from which information can be obtained in the survey. The target population is also known as the scope of the survey and the survey population […] as the coverage of the survey. [From: SDMX MCV ver. 1.0] · Age, nationality, and residence commonly help to delineate a given universe, but any number of factors may be involved, such as age limits, sex, marital status, race, ethnic group, nationality, income, veteran status, criminal convictions, etc. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, etc. [From: DDI ver. 2.0]
MC02	Geographical coverage · Information on the geographic coverage of the data. Include the total geographic scope of the data (the largest geographic unit that is covered by the data), and any additional levels of geographic coding provided in the variables. [From: DDI ver. 2.0]
MC03	Observation unit · Observation units are those entities on which information is received and statistics are compiled. During the collection of data, this is the unit for which data is recorded. [From: SDMX MCV ver. 1.0] · Basic unit of analysis or observation that the entity describes: individuals, families/households, groups, institutions/organizations,

	administrative units, etc. [From: DDI ver. 2.0]
MC04	Classifications and standards applied · A classification is a set of discrete, exhaustive and mutually exclusive observations, which can be assigned to one or more variables to be measured in the collation and/or presentation of data. [From: SDMX MCV ver. 1.0]
MC05	Definitions of row/column labels; Definitions of variable concepts or of categories (including descriptions on inclusion/exclusion)
MC06	Measurement unit · Quantitative data is data expressing a certain quantity, amount or range. Usually, there are measurement units associated with the data, e.g. meters, in the case of the height of a person. [From: SDMX MCV ver. 1.0]
MC07	Time reference/period · In one sense, this is synonymous with base period (Base period is the period of time for which data used as the base of an index number, or other ratio, has been collected). It may also refer to the length of time, e.g. week or year, for which data is collected. Population, statistical units and variables relate to specific times, which may be limited to a reference time point (e.g. a specific day) or a reference period (e.g. a month, calendar year or fiscal year) [From: SDMX MCV ver. 1.0]
MC08	Regional units · Geographic unit. Lowest level of geographic aggregation covered by the data. [From: DDI ver. 2.0]
MC09	Comparability over time (break in series, missing data) · Some users, especially those working in the macroeconomic field, use time series rather than point estimates for a date. The stability of the concepts and methods of measurement is very important to them. Details of changes in definitions, coverage, or methods should be provided. [From: Europastat, 2000]
MC10	Specific precautions
MC11	Source of the data
MC12	Explanation of standard symbols in tables
MC13	Information on copyright, restrictions of usage
MC14	Contact points for additional information

Part II - To be filled out for each instance of metadata elements

1. Table ID
2. Metadata category
3. Item ID
   Item ID consists of three parts: Table ID + metadata category code + two digit number. The last two-digit number is necessary in order to distinguish different instances of the same category of metadata for the same table.
4. URL of the reference page:
   URL of the page containing the metadata. If the table page is the reference page (i.e. the metadata is in the same page with the table), enter "same with the table
5. Title of the reference page:
   Leave it blank if the table page is the reference page. If there is no title in the reference page (neither in the main text nor in html <title> element), enter "no title
7. Label of the metadata unit:
   If the reference page is the metadata unit, enter "same with the ref. page
8. URL of the metadata unit, if applicable.
   If there is an anchor for accessing to the metadata unit, record the URL including the anchor

9. Proximity to the table – The distance from the table page to the reference page. The distance is measured by the number of necessary navigational actions (i.e. following a link), and should be based on the shortest possible path[4].
   01 – In the same page
   02 – Direct link
   03 – Two links
   88 – Other

10. Granularity of the link – This measures the granularity of both ends of the link. For example, a link can be provided from a specific part being described (e.g. column heading) to a relevant text segment (a metadata unit) within a reference document (an anchored text link). In that case, the link is from the table part to the metadata unit.

A. Direct link
   01 – From the table part to the metadata unit
   02 – From the table part to the reference page
   03 – From the table page to the metadata unit
   04 – From the table page to the reference page
      In case the reference page is also the metadata unit, record it as 'to the reference page'
   88 – Other

---

[4] For example, in BLS site, the distance from CPI table 1 (http://www.bls.gov/news.release/cpi.t01.htm) to the FAQ page that contains "Statistical population item is 2. There are multiple paths from the table to the page. 1) [Table of Contents] (a link provided under the table to get to CPI table of contents page) → [CPI Home] → General Overview page (which links to FAQ page for "population groups) → FAQ page. (4 clicks); 2) [CPI Home] → FAQ (2 clicks)
In locating metadata items, PDF or "text version of entire news release is not considered as a source of metadata.

Record the link text: _____

B. Two links

01 – Out-link from the table & In-link to the metadata unit
02 – Out-link from the table part & In-link to the reference page
03 – Out-link from the table page & In-link to the metadata unit
04 – Out-link from the table page & In-link to the reference page
88 – Other

Record the link text

- Out-link: _____
- In-link: _____

11. Modularity of metadata unit – This part is to look at whether a description for a single metadata element is in a distinct and coherent text segment.

A. Is the reference page is the metadata unit? (Yes/No)

01 – Yes
02 – No

B. Size of the metadata unit

01 – One paragraph
02 – Two to three paragraphs
03 – More than three paragraphs
88 – Other

C. Coherence: What portion of the metadata unit is about the specific metadata element in question

01 – The whole unit is about the metadata element
02 – More than 50% of the unit is about
03 – More than 30%, less than 50% of the unit is about
04 – Less than 10%