THE REALITY OF METADATA: ARE COMMON DATA DEFINITIONS POSSIBLE?

Jacob Bournazian, Dwight French, Mary Ellen Golby, Perry Lindstrom, Renee Miller, Louis Schloss, Energy Information Administration

Keywords: metadata, data definitions

In discussions of metadata, the need for specific definitions often emerges. Usability tests of Web sites confirm this need. During such tests Energy Information Administration (EIA) staff members have found that users are often puzzled by our terminology [1]. While making a glossary of terms available sounds like a straightforward task, there are many issues involved. This paper discusses EIA's experience in reconciling multiple definitions of the same terms. This is one step (albeit a major one); other steps necessary for the development and implementation of a glossary are also discussed. This paper concludes with a discussion of lessons learned in the process.

EIA's data definitions were originally developed for individual surveys so that survey respondents would report data correctly. EIA is now expanding its focus to data users and wants to ensure that customers understand EIA data and use the data appropriately. We, therefore, want definitions that the users will understand but that will also be appropriate for the respondents. The goal is to reach a broad audience consisting of survey respondents, technically-oriented customers, the media, others of varying degrees of technical sophistication, international customers who will most likely reach us through the Web, and the general public.

How did multiple definitions of energy terms arise in the first place? The answer probably goes back to the beginning of EIA. When EIA was formed in 1977, the agency inherited data collections from the Bureau of Mines, the Federal Energy Administration, the Federal Power Commission, and other agencies. The data collection forms and their accompanying publications were developed at different times and places and for different purposes. Furthermore, analysts had different ideas as to what constitutes a definition.

Why have multiple definitions persisted? That answer probably goes back prior to the formation of EIA. There are different definitions and usages in other contexts as well. An example from a completely different context is that "marinara" sauce in Italy and the United States is not the same. While energy analysts are not necessarily separated physically by mountains and oceans as the U.S. and Italy are, they have their own barriers. Upstream analysts (those who are concerned with energy production and reserves) and downstream analysts (those who are concerned with energy sales and prices) developed their own dialects over the years. Prior to the Web, they interacted little so there wasn't much confusion. This has all changed with the Web. Data on upstream topics are now just a few mouse clicks away from data on downstream topics, so that different definitions can result in confusion.

In response to concerns about multiple data definitions, EIA chartered an official cross-organizational team, the Common Data Definitions Team (CDDT), in February of 1998. The general objective stated in the Team's charter provided that the Team insure that EIA has common and accurate data definitions in most cases. The Team agreed to develop an underlying set of principles to guide it in its consideration of the content and format of definitions. The following sections describe the principles and procedures the Team developed, challenges encountered, accomplishments, remaining problems, and lessons learned.

Principles and Procedures for Developing Common Definitions

Some of these principles seemed obvious from the start; others evolved over time as the Team encountered new types of reconciliation problems. The major principles that the Team followed were to:

- Group terms according to subject matter. As an approach to developing single definitions, the Team decided to group terms by the type of energy they are related to such as "Coal" or "Petroleum." This approach allowed the Team to study the consistency of the revised definitions with the definitions of related single-definition terms that would otherwise not have been under our purview.

- Begin definitions with a generic statement. Whenever possible, the Team began definitions with an overall generic statement that was intended to serve as common ground for data users. More specific information was provided as needed for an understanding of the terms.

- Limit supplementary descriptive information. The Team made every effort to limit definitions to the minimum amount of information required to uniquely define the terms. In some cases, additional information was considered helpful. In those cases, a supplemental section of the definition, beginning with the italicized word, *Note*, was provided in order to include this additional information. These notes covered such information as references to specific instructions to survey respondents, caveats about limitations of the data for data users, or specific information needed for a complete understanding of a definition.

- Minimize subject matter specificity. Definitions that applied across EIA offices or subject-matter areas were written in general language without program-specific references. For example, terms such as "spot price," "stocks," or "reserves" pertain to multiple energy sources and were defined by using general language. The only definitions that were written in terms of a specific subject matter were those that pertained to one energy area.

- Make definitions timeless. The Team eliminated information in a definition that could change over time.

- Make definitions internally consistent. The Team reviewed terms with single definitions when they were related to terms with multiple definitions to ensure consistency. For

example, in reconciling multiple definitions for terms pertaining to gasoline, the Team reviewed the definition for the term "octane," although this term only had one definition.

- Minimize use of technical terms and jargon. The Team strived not to use jargon and to limit the use of technical terms. When a technical term was used within an official EIA definition, the term was bolded and cross-referenced.

Although the Team consisted of representatives from across EIA, we realized we did not possess all the knowledge necessary to deal with the wide variety of subject matter contained in the multiple definitions. We, therefore, sought input as appropriate during the developmental as well as during the formal review process (detailed directly below) from: subject matter experts from EIA program offices, the Environmental Protection Agency, the U.S. Geological Survey, the American Petroleum Institute, and the International Energy Agency.

The Team followed a rigorous review process once definitions were drafted. When the Team approved a draft set of definitions, we sent the draft to all EIA staff members requesting comments, with two weeks allowed for reply. After the comment period, the Team reviewed all comments received, determined the changes we would make to the draft definitions as a result, documented our responses to all comments, and sent the revised definitions out to EIA staff a second time with a request for comments, this time with a one-week turnaround. After responding to this second round of comments, the Team announced that the definitions were final and were to be used in all EIA products. In some cases there were additional steps in the process, such as meetings with interested staff members.

Challenges Encountered in Creating Single Definitions

In most cases, the Team was able to develop single definitions that were technically accurate and that appropriately represented the interests of the organization(s) that had created the multiple definitions. However, in some instances the differences were irreconcilable. Four areas of irreconcilable differences were: energy-use sectors, crude oil, oxygenated gasoline, and the United States.

- Energy-Use Sectors. The most difficult instance of irreconcilable differences occurred in the case of the energy-use sector terms -- residential, commercial, industrial, transportation, and electric power. This has been a long-standing problem for EIA where different program areas have been using various concepts of these sectors because of differing user needs. Not surprisingly, everyone wanted to continue to use these terms in their survey forms and data products. Changing program concepts of the sectors to match a standard definition was perceived to be impractical. The perception was that there were limitations in how survey respondents who were energy suppliers could report data by energy-use sector and that the cost of surveying energy end-users (an alternative approach to collect the data) would be prohibitive. Using some kind of program-specific term (e.g., "residential natural gas consumption" in table headers rather than just "residential") would be too confusing and too costly to implement.

  Several discussions and meetings on this topic took place involving EIA staff, the American

Statistical Association Committee on Energy Statistics, and EIA Senior management. As a result, for this one set of terms the Team developed definitions which reflect the core of the definitions used by each of the program areas. However, these definitions do not exactly match the coverage or output of any EIA data program specific to any of the sectors. The Team then developed an "Energy-Use Sector Guide" that explained the variations in coverage among the EIA data programs that categorize information by sector. The guide appears in the "A-Z" listing on the EIA Web site and will be referenced in EIA's integrated publications.

- Crude Oil. EIA had seven definitions of crude oil in the 1995 "Glossary" [2], which served as the initial frame for the Team's work. Some definitions included "lease condensate" (a mixture consisting of pentanes and heavier hydrocarbons, recovered in lease separation facilities) and others did not. To conform to common usage, EIA Senior management advised the Team to define crude oil to include lease condensate and to create and define another term for crude oil excluding lease condensate (shown in the Reserves publication). The Team therefore developed definitions for "crude oil" and "crude oil (less lease condensate)."

- Oxygenated Gasoline. EIA had two different definitions of oxygenated gasoline: one for petroleum supply data and one for petroleum marketing data. Petroleum supply data measure the production of oxygenated gasoline and include all gasohol. Petroleum marketing data, on the other hand, measure the price of oxygenated gasoline and include only gasohol sold in carbon monoxide nonattainment areas. These differences were irreconcilable due to limitations on what respondents could report and a desire to satisfy diverse customer needs. In this situation, the Team decided to define two terms: "oxygenated gasoline" and "oxygenated gasoline (includes gasohol)." The first term includes data on gasohol that is intended for sale inside carbon monoxide nonattainment areas and the second term includes all gasohol.

- United States. While the "United States" is generally defined as "the 50 States and the District of Columbia," EIA data programs may include data from Territories and other political entities outside the 50 States and the District of Columbia, including Puerto Rico, the U.S. Virgin Islands, Guam, American Samoa, Johnston Atoll, Midway Islands, Wake Island, and the Northern Mariana Islands. For these programs, data products will contain notes explaining the extent of geographic coverage included under the term "United States."

Occasionally, different EIA programs defined very general terms differently, and the differences were due entirely to variability in survey program operations or specifications. In these cases, the Team either picked one definition as the one for the general term and proposed that the program-specific term names be used for the other definitions; or the Team advised that all terms be renamed and did not define the general term. In a few cases, two concepts of a term were associated and yet easily distinguishable from one another in the context of use. In these cases, the Team developed definitions that encompassed both concepts. For example, the term "production," represents both the process of extracting resources and, as a data category, a measure of the amount of energy resources extracted.

In addition, there were some terms, such as "demand," "consumption," and "production," that represent generic economic principles. Rather than define these terms, the Team developed energy-specific labels (e.g., "energy demand") and defined them. The Team also allowed program areas to retain their specific definitions by maintaining the program-specific terms that were already in existence, e.g., "production, crude oil."

Accomplishments

The Team wrote and revised definitions for 262 existing terms and wrote definitions for an additional 89 new terms.[1] The new definitions appear on the EIA Intranet site. During this process, the Team reviewed approximately 1,000 definitions. In addition, the Team wrote an "Energy-Use Sector Guide" for EIA's Web site to explain variations in the definitions and data coverage across program offices for terms relating to energy-use sectors. A link contained in the energy-use sector definition's file enables customers to easily refer to the "Guide" for summarizing differences in energy-use sector data published by each EIA program office for the residential, commercial, industrial, electric power, and transportation sectors.

The Team reviewed and responded to more than 200 comments from EIA staff before finalizing the new definitions. Each comment was addressed and, on several occasions, the Team's response generated additional comments from staff. The revised definitions represent a consensus among the EIA staff as well as the Team's effort over the past three years. To avoid the problem of multiple definitions in the future, the Team wrote procedures and a new EIA Standard for writing definitions for new terms or changing existing definitions. The standard was adopted in September of 1999.

Beyond Multiple Definitions: Remaining Problems

EIA now has a new set of definitions written in standard language and format to the maximum extent possible. The Team wrote the definitions for a Web-based glossary. The bolded terms shown within the definitions are links. The next step will be making the Web-based glossary a reality. In addition, there remains a large body of definitions that the Team has not addressed, because they were outside our purview. If EIA intends to make a complete glossary available to the public, then additional efforts will be required in order to put these other definitions in standard format. Some of the problems we encountered are as follows:

- Definitions that are not common or generic, but that are fuel specific. This is by far the biggest challenge to making EIA's glossary something that would be appropriate for distribution on the Web or in hard copy form for the general public. An example would be the term "capital cost." There is only one definition for capital cost in EIA's internal "Glossary." That definition reads, "…cost of mine development and mill or plant construction and the equipment required for the production of uranium from a property, excluding sunk cost." While this definition might be fine for the uranium industry, it is not broad enough to serve as a definition for "capital cost."

[1] 51 new terms were written for electric power surveys and 38 terms were written for greenhouse gases and related terms.

- Related terms that may or may not contain multiple definitions but that depend on a shared concept that is not consistently or completely defined. An example of this would be the terms related to a "spot market." The first term is "spot market (uranium)." The second term is "spot purchase" as it relates to sellers and importers of crude oil. The third term is "spot purchases" and it uses the general term "fuel." The only true generic term is "spot-market price," which is a term not specific to any particular fuel or form of energy.

- Multiple terms for the same concept. Since the purview of the committee is multiple definitions for the same term, we did not visit definitions that were different but that could mean the same thing in practice. An example is "Direct load control" versus "Direct electricity load control."

- Terms that are outside of EIA purview that may be included on survey instructions but that are not needed in an official EIA glossary. Examples would include scientific or statistical terms that are not directly energy-related. These terms are in the internal "EIA Glossary" but would most likely need to be culled before the glossary is published.

Lessons Learned

To address customer needs, the Team decided that all definitions should begin with a generic, non-technical description. Technical information would follow the generic definition. Where needed, a note at the end would clarify related concepts or the scope of the data collected. In completing its work, the Team has seen the question emerge over and over again about whether a definition that is suitable for general users will be suitable for survey respondents and vice versa. The Team has received comments both from technical people within EIA stating that the definitions did not have enough technical material and from researchers outside the agency stating that the definitions are lengthy and information rich [3]. The definitions are now in the process of being implemented. The real test will be comments from users (both internal and external).

Dividing the terms into groups and sending them out for review as they were completed worked well, as did the open review process itself. Everyone in EIA was given the opportunity to comment. All staff members, however, did not respond with comments on a timely basis. A public relations campaign upfront may have helped. One way of doing this would be making presentations at staff meetings of the different divisions and branches to inform them of what was going on and that timely comments were crucial.

In thinking about the end product, we have seen several issues emerge. One is that we have not addressed all terms, only those with multiple definitions. These terms will conform to the definitional protocol while the others will not. Some of the terms with one definition are closely related to terms with multiple definitions. The Team has considered how to treat these terms on a case-by-case basis. The main lesson we learned from this is that there are many obstacles to overcome in attaining common data definitions and in going from an internal document to something that is appropriate metadata for the general public. Furthermore, it is important to

keep the lines of communication open among the various definition developers to maintain common data definitions in the future.

References

1. Blessing, Colleen; Bradsher-Fredrick, Howard; Miller, Renee; Rutchik, Robert, and Ware-Martin, Antoinette (1999). "Cognitive Interviewing: Applications to Evaluating the EIA's Web Site." Proceedings of the Section on Government Statistics and Section and Social Statistics. American Statistical Association, pp 245-250.

2. Energy Information Administration (1995). "EIA Glossary of Energy and Energy-Related Terms and Definitions." Washington DC.

3. Digital Government Research Center, Columbia University and University of Southern California. Presentation June 2001 Reception on Capitol Hill.