Committee for the Coordination of Statistical Activities

Conference on Data Quality for International Organizations
Rome, Italy, 7 and 8 July 2008

Session 3: Dissemination platforms to make data more accessible and interpretable

Accessibility and clarity: The most neglected dimensions of quality?

Steven Vale, UNECE

Introduction

The dimensions of accessibility and clarity are included in most statistical quality frameworks[1]. They are typically listed alongside other dimensions such as relevance, accuracy and coherence, for which there are well established and detailed means of assessment. However, for accessibility and clarity, the means of assessment are usually rather less developed, or sometimes completely absent.

To meet the requirements for accessibility, it is often seen as sufficient to make data available via the internet, whereas clarity is seen as satisfactory if a few footnotes or links to definitions are provided. This paper argues that far more is needed to meet user requirements in these areas, recognizing that accessibility and clarity are key quality components as far as users of statistics are concerned, and that the concept of quality is basically about meeting user needs. If users can not easily access data in the format they need, or if they do not understand the associated metadata, the data have little real value, even if they are perfectly accurate and coherent.

Accessibility – What does it really mean?

A recent survey of the web sites of the fifty-six UNECE member countries[2] found that almost all disseminated at least some basic statistical data free of charge. However the formats in which those data were available, and the ease of finding particular data sets varied considerably from country to country. Around sixty percent of countries offered some sort of database where users could select and download data, often in several different formats, whereas the remaining forty percent only offered pre-defined tables in static formats such as pdf, html or Microsoft Word / Excel.

---

[1] The term “clarity” is sometimes replaced by “interpretability”, but the meanings are similar.
[2] See the presentation “Current Practices in Data Dissemination via the Internet” for the UNECE Workshop on Developing Data Dissemination Systems, May 2008: http://www.unece.org/stats/documents/2008/05/dissemination-systems/wp.4.e.ppt

To make data truly accessible to a wide range of users it is necessary to consider several different formats and several different types of user. A simple categorization of users into the following three groups is becoming increasingly popular[3]:

- ***Tourists*** – these are novice or infrequent users, and typically make up the majority of individual users. They are looking for basic data either out of curiosity, or to inform personal decisions. They want to be able to find and view data quickly and easily, they prefer low levels of complexity and need only limited functionality.
- ***Harvesters*** – these are intermediate and fairly frequent users, who are looking for data to inform basic research or economic decisions. They will accept increased complexity if it results in addition functionality and flexibility in the way they can view and download data.
- ***Miners*** – these are expert users, typically small in number, but using large volumes of data on a regular basis, often for detailed research or analysis. They want high levels of functionality and flexibility, and are willing to invest some time to learn how to use a data interface.

"Tourists" are often happy with data in static formats, as long as they are easy to find and interpret. Therefore quality assessments for this group of users should focus more on ease of access and search, as well as logical and clear presentation of data. In the international context, they are the most linguistically demanding, as they prefer to find data in a language with which they are familiar, rather than to spend time trying to translate statistical terminology.

"Harvesters" and "miners", however, have rather different needs, they prefer the database approach to statistical dissemination, where they can select and download just those data that are of interest. There is still, however, a tension between these groups, as "harvesters" prefer relatively simple interfaces with limited functionality, whereas "miners" want to be able to do much more, and are typically more expert in using statistical databases. Once they have found the data they are looking for, both groups want to extract and download sub-sets in a variety of formats.

"Miners" may also be interested in integrating different data sets, and performing statistical analyses, though given the range and complexity of software available to do this sort of work off-line, increasing amounts of which are now freely available under open-source licenses, it is debatable whether there is always sufficient value-added to justify the development and implementation of this sort of functionality in statistical database interfaces.

It is therefore clear that to meet the needs of users in terms of accessibility, a statistical agency should offer data in several different formats, whilst balancing this with the requirement of not confusing the user. Options such as "view key figures" and "explore detailed data" can help guide different users to the most appropriate formats.

---

[3] For example, see: http://www.unece.org/stats/documents/2008/05/dissemination-census/wp.3.e.pdf. A similar approach, with different category names can also be found in an earlier paper from Statistics New Zealand: http://www.stats.govt.nz/census/about-2006-census/methodology-papers/developing-census-product-service-mix.htm

One important fact, often overlooked when designing statistical data interfaces for the web, is that users are rarely tied to a single source; they can and do use a mixture of national and international, official and non-official sources. If, as is typically the case, the web interface for each source is designed in isolation, users have to spend valuable time becoming familiar with new layouts and terminology every time they visit a new site. This is a serious impediment to accessibility when we take a holistic view of official statistics.

By improving the standardization of data interfaces, the official statistical community could make concrete improvements to accessibility. A good starting point would be to introduce a single classification of statistical domains, so that users would always know where to find, for example, data on employment, rather than having to learn the classifications used by each organization. Such classifications are often rather producer-oriented, driven by internal organizational structures, which may not be the most logical or accessible basis for classifying data as far as users are concerned. The classification of statistical subject-matter domains used for the Database of International Statistical Activities[4] is being adopted as a standard within the SDMX[5] initiative, and would therefore form a good basis for such harmonization. Other harmonization actions focusing on areas such as terminology, presentation and the appearance of interfaces could also be envisaged.

Traditionally accessibility has focused on passive dissemination, i.e. making data available in the hope that someone will use them. The extent to which more pro-active approaches to dissemination, such as marketing exercises targeting actual and potential users, could improve the accessibility dimension of quality, is perhaps open to debate. However, it is clear that making users aware of the existence of data will certainly not diminish accessibility, so it would seem appropriate that assessments of accessibility take into account pro-active dissemination measures.

Accessibility and Visualization

The previous section considered accessibility mainly in terms of tabular data, but there is a rapidly growing interest in finding new ways to present data graphically. The data visualization techniques used or offered are therefore also relevant when assessing accessibility. The old saying that a picture is worth a thousand words can certainly be true for good visualizations, but there should be a clear distinction between “ready-made” visualizations provided by data suppliers, and “self-service” visualizations produced by users.

There are many different data visualization techniques, all of which have their limitations, and some of which are only really useful for very specific types of data. Visualizations provided by data suppliers can be matched to the data they illustrate, and can be produced according to agreed guidelines or in line with best practices. “Self-service” visualizations are, however, virtually impossible to control, particularly when users are presented with a large menu of different options. It is very easy for less experienced users, or those that are simply in a hurry, to produce poor quality, or even

[4] See: http://unece.unog.ch/disa/
[5] Statistical Data and Metadata eXchange – www.sdmx.org

misleading visualizations. There is also a risk that the true message behind the data could be obscured or distorted in an attempt to produce something more visually appealing. Of course, deliberate distortion to advocate a particular view is also an ever-present risk. In this sense, it could be argued that giving users a choice of overly-complex visualization techniques might even be detrimental to both accessibility and clarity.

In any case, more advanced users such as the "miners" described above, are likely to have access to a wide range of external data visualization tools after they have downloaded the data. This means that attempts to improve accessibility by providing advanced users with too many graphic visualization tools, or tools that are too complex, could be a waste of resources, particularly if this becomes some sort of competition between statistical agencies. Using resources to provide good "ready-made" graphics that communicate key messages to "tourists" and "harvesters" is likely to be much more cost-effective in terms of improving accessibility.

Clarity – We could do better!

Turning to the concept of clarity, this is also an area for which a coordinated reflection is needed on what the users really want. Significant advances have been made over the past ten years in making statistical metadata available to the users of our data. Most organizations now have some sort of on-line glossary, and many provide links to electronic versions of methodological manuals and texts. However, these documents have not, for the most part, been sufficiently adapted to the purpose of providing information for the majority of data users. They tend to be written in a way that is only really understandable to subject-matter experts, and are therefore not sufficiently user-friendly for the "tourists".

As an example, the term "actual individual consumption" is defined in several on-line statistical glossaries as follows:

> *"Actual individual consumption is measured by the total value of household final consumption expenditure, non-profit institutions serving households (NPISH) final consumption expenditure and government expenditure on individual consumption goods and services."*

The original source of this definition is the SNA 1993[6], a manual written by National Accounts experts, for National Accounts practitioners, not for "tourists". It is therefore not surprising that such a definition is not very clear to these users. In addition to the standard definitions of statistical concepts, explanations and interpretations designed for "tourists" should also be provided[7]. This would not only help to improve clarity, but could also help to build statistical literacy amongst all users. Sharing the task of providing these explanations between international organizations would not only be the most efficient way to proceed, but would also help to improve standardization.

[6] System of National Accounts, 1993 version, see: http://unstats.un.org/unsd/sna1993/toctop.asp
[7] It should be noted that the OECD have moved at least partly in this direction by providing information on context in their on-line glossary. See: http://stats.oecd.org/glossary/detail.asp?ID=36

Some years ago, Eurostat and the OECD produced "pyramid" models of metadata[8], to show how the differing requirements for metadata from different user groups could be met using a "drill-down" approach, where the users are given the option to move between metadata of different levels of complexity according to their needs. This approach is also relevant to setting priorities for improving clarity, in that the degree of clarity required for "tourists" is greatest for the higher level metadata such as table headings, footnotes, concepts and definitions. However, few "tourists" are ever likely to consult detailed methodological manuals, so the "translation" of these for non-specialists is less of a priority.

Another application of this principle could concern metadata on quality. At the highest level, a simple overall quality indicator could be presented, e.g. using the "traffic light" approach: green for good quality, orange for caution, and red for danger. This, of course, should be supplemented by links to more detailed reports for specialist users.

Conclusions

If international statistical agencies are serious about improving the accessibility and clarity dimensions of data quality, there is a need to engage more with specialists in areas such as communication and education to help us get our messages across, particularly to our largest actual and potential user group, the "tourists". We need to find a way to simplify our outputs for these users whilst retaining the key messages, and not alienating our more advanced users, the "harvesters" and the "miners". The best way to do this seems to be to offer different formats and different levels of detail for both data and metadata.

There is a clear role for international organizations to lead the way in standardizing the user experience on statistical web sites. If we can agree standards amongst ourselves, it will be easier to persuade national statistical organizations to adopt them in the future. Above all, we need to think more about the "tourists" when we are evaluating the accessibility and clarity of our statistics. If we do, we will see that there remains considerable scope for improvement.

[8] See "The Metadata Problem in a European Context", S. Vale and M. Pellegrino, Eurostat, 2000. www.unece.org/stats/documents/2000/11/metis/crp.1.e.pdf, and Section 1.4 of Main Economic Indicators, Comparative Methodological Analysis: Consumer and Producer Price Indices, OECD, 2002. www.oecd.org/dataoecd/60/61/1947731.pdf