Chapter LIII
Statistical Dissemination Systems and the Web

Sindoni Giuseppe
Eurostat, Luxembourg

Tininini Leonardo
CNR - Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti", Italy

INTRODUCTION

The Web is increasingly used as a preferred medium for A2C (administration to citizens) and A2B (administration to business) service delivery. An increasing number of government initiatives are aimed at making access to electronic records easier for the general public. For example, the Electronic Record Archives program of the U.S. National Archives and Records Administration is aimed at preserving virtually any kind of electronic record, free from dependence on any specific hardware or software, and at enabling customers to find records they want and to deliver those records in formats suited to customers' needs (Lake, 2006). This in particular will include records coming from the 2010 U.S. census. International professional associations are increasingly paying attention to public availability of statistical data. For example, the last meeting of the International Association for Social Science Information Services & Technology (IASSIST, 2006) dedicated entire sessions to problems like knowledge and resource discovery, innovative data dissemination systems, and data-intensive Web site design.

In this context, the world of public statistics has the opportunity to exploit many efficient, flexible new technologies and standards to deliver better quality statistics in a more timely and accessible way. First, Web technologies can greatly reduce the gap between data producers and users as published data can be made available immediately after their production to a worldwide community of users. Second, as published statistical data are multidimensional, it would seem that the consolidated methodologies and techniques for data warehouse (DWH) modeling and navigation could be easily applied to support user navigation among the available data. However, this methodological and technological scenario cannot be implemented straightforwardly due to some specific features of statistical data in contrast with conventional business data, which require the introduction of specifically designed models and techniques.

In this chapter we review the main concepts at the basis of multidimensional (data warehouse) modeling and navigation. We also illustrate some

peculiarities of statistical data that make the implementation of a statistical data warehouse, that is, a statistical dissemination system enabling the user to perform a multidimensional navigation, a challenging issue in many aspects. Finally, we analyze the main characteristics of some of the most important systems for the dissemination of statistical data on the Web, distinguishing two main approaches, the former based on a free navigation of specific subcubes, and the latter on a constrained navigation of a single data cube.

BACKGROUND

In this section we review the main terms and concepts related to data warehousing and statistical databases (SDBs). A key concept common to both contexts is that of (statistical) aggregate data. These are obtained by applying simple aggregation functions (Cabibbo & Torlone, 1999; Klug, 1982), like the standard count, sum, min, max, and avg, or more complex statistical analysis functions to groups of elementary data (usually called microdata in statistical terminology). In statistical surveys, microdata are commonly obtained from questionnaires, but data extracted from public registries are also becoming increasingly important.

Microdata are rarely processed in their original form, but are instead transformed by specifically designed tools for extraction, transformation, and loading (ETL), performing a general reconciliation of the source data and a reclassification of some attributes, as well as some preaggregations. The resulting data are stored in the so-called fact tables (see Figure 1), comprising both dimension codes, used to classify the data, and measures, on which the aggregation and statistical functions are applied (Cabibbo & Torlone, 1997; Kimball, 1996). Each fact table is typically linked to a collection of dimension tables to form the so-called star and snowflake schemas (see Figure 2). Dimension tables are used to both decode the dimension codes and define groups of codes, which represent the levels of detail of the classifying dimensions and consequently also of the aggregate data. These levels are commonly known as dimension levels and are organized in dimension hierarchies (Jagadish, Lakshmanan, & Srivastava, 1999). For example, if D_i is an area dimension, $D_{i,1}$, $D_{i,2}$, and $D_{i,3}$ may correspond to the national, regional, and municipality levels, and may also have a temporal evolution that affects the corresponding aggregate data (Tininini, Paolucci, Sindoni, & De Francisci, 2002). Figure 2 shows an example of star schema with three dimension tables, one of which (the area dimension) has three dimension levels.

Figure 1. From microdata to aggregate data