Chapter 129
Automatic Metadata Generation for Geospatial Resource Discovery

Miguel-Angel Manso-Callejo
Universidad Politécnica de Madrid, Spain

Arturo Beltran Fonollosa
Universitat Jaume I de Castellón, Spain

ABSTRACT

Metadata are structured sets of data that describe other data and whose purpose is to improve knowledge about described resources. Metadata help to answer questions, locate, and facilitate the use of resources, but in addition to these features several authors propose another purpose: Make interoperability possible in a distributed environment such as Spatial Data Infrastructures (SDI). Metadata also have been classified according to their nature, the way they are stored together with the resources, or how to obtain them. Metadata must be created to respond to current needs, especially resource discovery, and anticipate future needs based on interoperability. Several referenced authors in this domain have identified different ways of creating metadata: by editing, by extraction, by calculation, or by inference. Most of them are in favor of automating metadata production to avoid mistypes and interpretation errors, and to prevent creators from becoming discouraged by this monotonous work.

In the metadata generation context, metadata extraction is the first and most important stage in the production chain and has an enormous complexity due to the huge variety of storage formats for geospatial datasets. In addition, the authors analyze the current situation and importance of metadata in information systems and particularly in SDI. This chapter identifies and justifies the need to automate the metadata generation. In this context, the different metadata points of view according to their functions and interoperability levels are analyzed. Afterwards, different metadata generation methods and workflows, and various metadata generation related tools are reviewed, respectively. Finally, the authors introduce topics related to the automatic metadata generation that have neither been studied in depth nor prototypically implemented as future works.

DOI: 10.4018/978-1-4666-2038-4.ch129

INTRODUCTION

The concept of metadata is hardly new—the most common definition of the term *metadata* is "*data about data*," with the first references to this term appearing in the context of geographic information, in ANZLIC (1996) and Kildow (1996). If we look for the origins of the term *metadata*, we will find its roots in the Greek word "μετα," "beyond" and the word "data," the plural of the Latin term *datum-i*, "piece of information" (RAE, 2011[1]). Therefore, the meaning of the word may be explained as "beyond data." However, according to Howe (1993), the term *metadata* did not appear in print until 1973, despite having been coined by Lack Myers in the 1960s in order to describe sets of data and products. In the literature related to this subject we find a good number of authors who provide the interpretation and scope of the practical and theoretical meaning of the term. Among these, we find Caplan (1995), Milstead and Feldman (1999), Ercegovac (1999), Sheldon (2001), and Steinacker *et al.* (2001), Swick (2002), and Duval *et al.* (2002), or Woodley *et al.* (2003). Summing up the contributions of all these authors, we may define the term eclectically as *the structured set of data that describe other data and whose purpose is to improve our knowledge of the described information and help us answer such questions as 'what,' 'who,' 'where,' 'when,' 'how much,' and 'how.'* They may also be described as those autonomous products that, linked to the data, allow us to keep an inventory of these, enabling its publication and reference value through the catalogues kept in SDI and, finally, allowing for the reutilization of data. The importance of metadata has been recognized by entities such as the EU's INSPIRE[2] Directive, and also by the endorsements of the GSDI[3] initiative.

Moreover, Caplan (1995) acknowledges that the concept of metadata is used to avoid the prejudices developed by professionals in the field of information, who are closer than most to the world of libraries: computer technicians, software designers, and system engineers. Finally, metadata are used to describe the context, the quality, the condition or the characteristics of the data (Milstead & Feldman, 1999; Howe, 2003) in such a way that users can discover and understand their data sets, particularly in the context of Geographic Information (GI). For Zeigler *et al.* (2006), metadata is "*a hierarchical concept in which metadata are a descriptive abstraction above the data it describes."*

Various experts are in favour of assigning the task of metadata creation to the owners of the geospatial datasets (geodata), in the belief that these owners are best suited to provide information about their data (Greenberg, 2004; Kolodney & Beard, 1996). In practice, metadata creation has occupied a secondary role within organizations, having been created after its production. For this reason, some organizations have considered the creation of metadata as an additional cost (Najar, 2006). This fact has been criticized by several studies; for example, in the CGIAR-CSI (2004) study we find the following statement: "*The creation of metadata to novel data producers might seem burdensome, but the long term advantages are far superior to the disadvantages of the initial burden of implementing a Metadata policy within an organization. The initial expense of documenting data clearly outweighs the potential costs of duplicated or redundant data generation."*

One natural consequence of the fact that metadata creation does not occur simultaneously when the actual geodata is compiled, is the presence of errors, which sometimes turns the creation of metadata into an almost impossible task (Kolodney & Beard, 1996; Caplan, 2003; Leiden, et al., 2001). Moreover, the standards are complicated and extensive. For instance, standard ISO19115 (2003) defines more than 400 elements for metadata. Consequently, manual creation of metadata is a monotonous, harsh, resource-consuming and is prone to contain errors. As Manso and Bernabe (2009) show as conclusions of the study of "Characterization of Temporal Cost and Error Type and