XML-based Metadata Management for Biological Data

J. L. Campos dos Santos [1,2] and R. A. de By [1]

1. International Institute for Geo-Information Science and Earth Observation
Hengelosestraat 99, P.O. Box 6
7500 AA Enschede – The Netherlands
Emails: {santos, deby}@itc.nl
Tel: +31 (0)53 487-4374, Fax: +31 (0)53 487-4335

2. The National Institute for Amazon Research
André Araújo Avenue 2936 – Petropólis
69.083-000 - Manaus – Amazonas - Brazil
Emails: lcampos@inpa.gov.br
Tel: +55 (0)92 643-3032, Fax: +55 (0)92 643-3311

Keywords: Metadata, Schema, XML, Web Environment, Biological Data

ABSTRACT

For more than a century, Amazonian biological data have been collected, primarily by single or small groups of researchers in small areas over relatively short periods of time. The world is questioning ecologists "on how ecological pattern and process vary in time and space, and what are the causes and consequences of this variability". For such questions to be properly answered, far more data are required than could feasibly be collected, managed, and analysed under the auspices of an individual or group of researchers. The strategy utilized to know the answer in a certain extent is to use information that has been collected and documented by other research teams.

A common practice among scientific interdisciplinary teams is to share data with expert colleagues to address specific questions. Biological data sets are neither perfect nor intuitive. The data which are collected recently, the sharing mode is kept in a close range to the data producer (same institutions or projects), whose know the subject and few additional information is needed, in order to data set be used and interpreted. Science teams outside of the specific subject area need highly detailed documentation to accurately interpret and analyse historic or long-term data sets, as well as, data from complex experiments.

Usually, researchers refer to their data as raw data, which are structured in rows and columns of numeric or encoded sampling observations. The usefulness of such data can only be assessed when they are associated to either a theoretical or conceptual model. This requires the understanding the type of variable, the units adopted, potential biases in the measurement, sampling methodology, and a series of facts that are not represented in the raw data, but rather in the metadata. Data and metadata combined within a conceptual framework produces the so needed information. Additionally, information can be lost through degradation of the raw data or lack of metadata. The loss of metadata can occur throughout the period of data collection and the rate of loss can increase after the results of the research have been published or the experiment ends. Specific details are most likely to be lost due to the abandonment of data forms and field notes. Metadata will ensure to data users the ability to locate and understand data through time.

This paper presents a XML solution for biological profile metadata management on the Web environment. The use of XML schema for the definition of well-formed bio-metadata file's templates, allows the representation of descriptions of the FGDC Metadata Standard, which includes the Biological Data Profile. The XML metadata files can be deployed together with a XML Editor. The editor upload the XML file and allow users to insert the metadata information. After this process, the biological metadata can be submitted

via Internet for certification and stored in a XML repository. The repository accepts any number of well-formed XML metadata and maintains a single data representation of all the files it receives. The metadata can be retrieved, updated, or removed from the repository. Once the repository is indexed, search and query are available. This solution is in test at the National Institute for Amazon Research within the Biological Collection Program.