Chapter 14
Adaptive Acquisition of VGI to Fill Out Gaps in Biological Observation Metadata

Daniel Cintra Cugler
Federal Institute of Triangulo Mineiro, Brazil

Claudia Bauzer Medeiros
University of Campinas, Brazil

ABSTRACT

Biological observation databases store data on species observations, being used in many kinds of research. Such observations are often queried and/or correlated primarily using metadata parameters (e.g., spatial queries on metadata concerning regions where observations were performed). However, metadata are often missing - either blank attributes, or lack of metadata records - which hampers the use of the observations databases. Filling these gaps is challenging because metadata requirements change as researchers acquire new knowledge about their problems. Related work is limited because it does not take this knowledge evolution into consideration. This chapter presents an approach to acquire missing metadata records, which fully supports dynamic on-the-fly evolution of metadata requirements. As proof of concept, we implemented a configurable software platform to collect data from "human sensors" and other sensors. Among its many dynamic characteristics, it allows deployment of context-sensitive forms to be filled by volunteers, according to a location and a research target.

INTRODUCTION AND MOTIVATION

Biological observation databases contain priceless information that can be used to provide knowledge for broad kinds of research, e.g., global warming, species behavior, food production, etc. Such information can be acquired in many distinct ways and from many distinct sources. Observation metadata (the data that describe observations) play a major role in scientific research involving biological observations; metadata are the primary means to computationally process, correlate, and access observational data.

As such, they are considered by scientists as "first-class citizens", being often indissociate from the observations themselves. Such metadata may differ according to the domain, but the most important fields are supported by all kinds of observations -- i.e., "what" (what was observed), "when", "where", "how" (methodology) and "who" (observer). We point out that, throughout this chapter, the main assumption is that research that uses biological observations is highly dependent on the "where" metadata attributes – i.e., though not always explicit, processing geolocation information is a key issue here. In fact, "where" has become one of the most important pieces of information in observations, since a growing spectrum of studies are sensitive to geospatial information. In many cases, observation data and metadata are intimately associated, so that the difference becomes fuzzy – see remarks in Section "*Background and Related Work*".

In the past, observations were made mainly by domain experts, who needed to spend from days to months traveling and collecting information in field trips. Such observations were usually made manually, and metadata were annotated in paper, in natural language, most of the time following no predefined data structure.

The way of collecting and persisting biological observations and metadata has changed drastically – and not only due to technological progress (e.g., use of GPS-enabled devices to automatically capture the "where"), but also because new paradigms have appeared for data capture methodologies. First, an increasing number of observations started to be provided not only by domain experts involved in projects, but also by non-expert volunteers. Such information, named VGI (Irwin, 1995; Goodchild, 2007) (Volunteered Geographic Information -- See Section "*Volunteered Geographic Information (VGI)*"), has introduced a new paradigm in the way of collecting data. Second, computational systems with predefined data structures are now commonly used to persist observational metadata. Third, different computational tools are being employed to acquire information, such as web-based forms and mobile applications. In spite of those changes, one feature remains the same -- there is often missing information in the metadata, which demands time- and cost-intensive curation processes.

Such missing information (here called *information gaps*) can bias research into wrong scientific conclusions or even hamper the use of the observations. There are countless reasons for gaps in biological observation metadata -- e.g., lack of equipment to measure environmental variables or lack of standards. The approach presented in this chapter generalizes and classifies such gaps into two groups: (a) incomplete information in the metadata records and (b) insufficient samples for some specific scenario.

- **Incomplete Information:** Here, not all record fields are filled out. For example, when recording animal vocalizations, identification of species is a hard task since many times the scientists do not see the animals whose vocalization they are recording. For this reason, metadata may, for instance, lack an animal scientific name, providing only partial taxonomic classification. Additional gaps may be caused, for instance, by lack of appropriate instruments at the observation site, or absence of adequate conditions to perform some measurement. This is a widely studied problem, for which the majority of solutions are based on data curation algorithms.
- **Insufficient Number of Samples:** Here, the problem is having an insufficient number of observation (and thus metadata) records that fit a scenario. This may occur, for instance, when a team of scientists wants to reuse data collected by another team, but the teams have distinct spatial sampling methodologies. This chapter is concerned with handling this second kind of gap, since, as we show, it is much more likely to profit from VGI as the other kind of gap. It must be pointed

Related Content

A New Image Distortion Measure Based on Natural Scene Statistics Modeling
Abdelkaher Ait Abdelouahad, Mohammed El Hassouni, Hocine Cherifiand Driss Aboutajdine (2013).
Geographic Information Systems: Concepts, Methodologies, Tools, and Applications (pp. 616-630).
www.irma-international.org/chapter/new-image-distortion-measure-based/70465

Estimating Urban Tree Metrics Using Terrestrial LiDAR Scanning
Tyler Jones, Luke Marzenand Art Chappelka (2022). *International Journal of Applied Geospatial Research
(pp. 1-10).*
www.irma-international.org/article/estimating-urban-tree-metrics-using-terrestrial-lidar-scanning/302092

Perspectives on Applications of Geospatial Technology and Landscape Ecology for
Conservation Planning in the Global South
Henry N. N. Bulley, Oludunsin T. Arodudu, Esther A. Obonyo, Aniko Polo-Akpisso, Esther Shupel
Ibrahimand Yazidhi Bamutaze (2023). *International Journal of Applied Geospatial Research (pp. 1-23).*
www.irma-international.org/article/perspectives-on-applications-of-geospatial-technology-and-landscape-ecology-for-conservation-planning-in-the-global-south/313941

Geospatial and Spatio-Temporal Analysis in Health Research: GIS in Health
Dimitra I. Sifaki-Pistolla, Georgia D. Pistolla, Vasiliki-Eirini Chatzeaand Nikolaos Tzanakis (2017).
Handbook of Research on Geographic Information Systems Applications and Advancements (pp. 466-487).
www.irma-international.org/chapter/geospatial-and-spatio-temporal-analysis-in-health-research/170001

A Geographical Analysis of Socioeconomic and Ideological Drivers of Hate Crime in the United
States
Richard M. Medina, Emily A. Nicolosi, Simon Brewerand Erin Moore (2021). *International Journal of
Applied Geospatial Research (pp. 39-56).*
www.irma-international.org/article/a-geographical-analysis-of-socioeconomic-and-ideological-drivers-of-hate-crime-in-the-united-states/273690