Semantic Linking of Data for the Brazilian Amazon Rainforest Statistics

Giovana Mira de Espindola[1], Tomi Kauppinen[2]

[1]National Institute for Space Research (INPE) – Brazil
giovana@dpi.inpe.br

[2]Institute for Geoinformatics (ifgi) – University of Münster – Germany
tomi.kauppinen@uni-muenster.de

Abstract. Different techniques, including remote sensing, are used to collect data about the deforestation in the Brazilian Amazon rainforest. However, a key problem for integrating all the resulting datasets for statistical analysis is the semantic diversity among these datasets. In this paper we identify key semantic interoperability challenges for the Brazilian Amazon rainforest statistics.

1 INTRODUCTION

The Amazon rainforest is a tropical moist broadleaf forest settled in much of northern South America, mostly in northern Brazil (BRITANNICA, 2008). The process of human occupation in the Brazilian Amazon was spatially and timely diverse and was largely induced by government policies and subsidies (ANGELSEN, 1997; MACHADO, 1998; BECKER, 2005).

Nowadays, recent deforestation in this region is mainly related to private investments in agricultural expansion, associated with large-scale cattle ranching, small-scale familiar farming and soybeans expansion. From 1988 to 2009, the total deforested area in the region varied from about 7,000 sq km to more than 29,000 sq km a year (INPE, 2009). The lower rates of deforestation since 2005 have been associated with control actions conducted by the Brazilian government, including law enforcement actions and the creation of protected areas, and partly to lower commodities prices in the international market.

Despite the decrease since 2005, current deforestation rates are still serious, raising the need for frequent and accurate assessment of forest loss. Remote sensing can be used to reveal regions where deforestation has taken place. Several existing data sets of heterogeneous temporal and spatial accuracy are maintained to document the results of remote sensing. A few countries have institutions to monitor changes in forest cover that have been in place for several decades, most notably Brazil (INPE, 2009).

The National Institute for Space Research (INPE) has four information systems for monitoring deforestation in the Brazilian Amazon: PRODES,

DETER, DEGRAD and QUEIMADAS. These systems are complementary and were designed to meet different goals, and in this article we will focus on PRODES, DETER and DEGRAD. Firstly, PRODES (Amazon Deforestation Monitoring Project) is a high-spatial resolution map of deforestation produced annually using remote sensing imagery. While PRODES captures the spatial detail required to generate area estimates of deforestation, there is some latency in creating the annual update products (HANSEN *et al.*, 2008). Secondly, to improve timeliness and meet the needs of other users, INPE has also incorporated the use of remote sensing imagery for forest change monitoring through the DETER (Near Real Time Deforestation Detection System) project (SHIMABUKURO *et al.*, 2007). DETER data are acquired daily and allow for the near-real time detection of large-scale change events. The DETER products enable quick responses by government agencies in enforcing forest land use policies. In summary, current INPE applications employ DETER for identification of new change hotspots in near-real time and PRODES for the precise areal quantification of change. Thirdly, INPE also maintains an innovative system (DEGRAD) to identify areas in process of deforestation. Being in charge of registering the partial fall of the forest, caused by fire or selective wood extraction, DEGRAD can be an important subsidy to the surveillance departments and this way can prevent the shallow cut phase.

Even with advances in monitoring deforestation, significant challenges remain including the need to understand the different trajectories of change in the region, which are translated into diverse rates and patterns of deforestation. Another challenge is the integration of all these data sets, which is not straightforward for several reasons. First of all, there is often a lack of syntactic interoperability. But even more crucial problem is the lack of semantic interoperability (KUHN, 2005) between data sets regarding tropical forests. Essentially, various datasets maintain information about semantically different phases of the whole process of deforestation, and there is semantic heterogeneity among the datasets.

In this paper we discuss and identify semantic interoperability issues that call for semantic integration of the datasets of deforestation, and to enrich them with formal semantics. This work includes development and application of ontological structures and rules for linking data, and publishing of results as linked open data[1]. We believe that these integrations could be useful to reason about the different trajectories of change in the region. We envision a combination of statistical and ontological models for that.

[1] http://linkeddata.org

2 SEMANTIC INTEGRATION OF THE BRAZILIAN AMAZON DATA

Deforestation is a process of change that can be described as steps. For the purpose of this article we will consider the deforestation process showed in Figure 01, where INPE's information systems are giving us data for two of our steps – 'Degraded Forest' and 'Deforested Forest'.

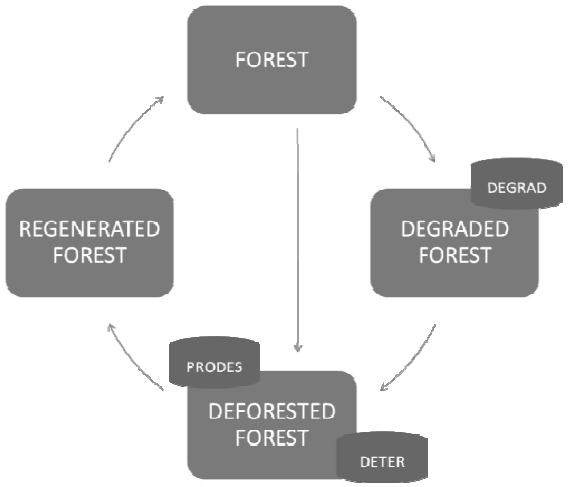

Figure 01 – The deforestation process and INPE's information systems.

The process of deforestation can vary in space and time, and statistical analysis of the varied data sets related to the Brazilian Amazon rainforest aims to provide better understanding of land change in the region. More precisely, interesting patterns found out in the data may potentially be used to predict changes in the area, and thus help to create better public policies. For example, cases of small versus large areas of deforestation can be associated with specific patterns (corridor, geometric, etc.) and actors (small farmers, large farmers, etc.) of change. However, the problem in conducting statistical analysis is that datasets often lack semantic interoperability, either between datasets or even within one dataset.

An ontology aims to enable semantic interoperability by capturing essential concepts that can be found either implicitly or explicitly in datasets. The ontology captures also rules to relate these concepts. In other words, the ontology is used to create relationships between data from different datasets, and hence make datasets semantically interoperable. Given the semantically integrated datasets, the hypothesis is that a statistical analysis will be more accurate and have a wider coverage, and hence results of it are more useful. For example, if all the concepts having different labels but same meanings are identified and modeled by ontology then statistical analysis can be conducted with all datasets sharing references to these concepts rather than using just a single data set. The goal of combining statistical approaches with ontological approaches is to be able to answer ques-

tions like: “What are the largest degraded areas among all the deforested areas?”, “How is a small area of deforestation evolving in time?”, “What is a small versus large area of deforestation?”. Here, an ontological approach will take care of semantic integration of data sets while the statistical approach will take care of finding interesting patterns or ranking results.

In the following we provide an analysis of different types of semantic interoperability issues regarding the Brazilian Amazon datasets.

- Concept-level mismatches
    - Closely related or hierarchical concepts. For example, different phases of land use are maintained in different datasets, e.g. one dataset is about forest degradation (DEGRAD) and another about deforestation (DETER, PRODES).
    - Concepts or attributes in two datasets have different names, but essentially the same meaning. For example, SCENE_ID (DETER) and PATHOROW (PRODES) both mean the remote sensing imagery identification. Another example is related to attributes for times, e.g. JULDAY (PRODES) vs. DIA, MÊS, ANO (DETER) which both mean a specific date.
    - Concepts in two datasets have same names but different meanings. For example, CLASS_NAME in DETER and PRODES refer to different sets of classes. Another example is TIME attribute in PRODES which means the year when some region was deforested, while TIME attribute in DETER means the date when some deforested area was identified.
- Instance-level mismatches
    - Instances have the same meaning (i.e. the extension) but different labels, e.g. a place name in different languages, and according to different spelling variations.
    - Instances have different meanings but the same name, e.g. two different places sharing the same name.
- Temporal mismatches
    - Temporal resolution differs, e.g. one piece of data is recorded yearly (PRODES, DEGRAD), and another data monthly or daily (DETER).
    - Meaning of time may differ in many ways, e.g. data with a timestamp “2010” could mean:
        - The average of daily activity during that year (e.g. deforestation).
        - The maximum of that year.
        - The minimum of that year.
        - The cumulative of that year.
- Spatial and spatio-temporal mismatches

- Differences in spatial accuracy, e.g. AREA attribute in PRODES has a better spatial accuracy than AREA attribute in DETER.
- Spatial resolution differs: one dataset might cover the municipality level while another covers a farm, a province, a regional, a state or even a country level. In addition, data might be from tessellations—e.g. polygons, cells, or triangles—of varied shapes and sizes.
- Differences in scales and generalization levels.
- Changes have happened in borders of places and in place names. These changes cause data to refer to places of different sizes, shapes, and names. Because the meaning of a place may differ over time, a statistical analysis might mix up with these different meanings.

3 CONCLUSIONS

In this article we identified several types of semantic heterogeneity that introduce challenges for conducting statistical analysis of the Brazilian Amazon rainforest datasets. We state that addressing the problem of semantic heterogeneity requires development of ontological structures for modeling essential concepts, attributes and units of datasets, and rules for reasoning about these structures to enable alignment of datasets. We envision that this will enable conducting wider covered and more accurate statistics of the Brazilian Amazon rainforest datasets.

Acknowledgements. The research has been partially funded through the International Research Training Group (IRTG) on Semantic Integration of Geospatial Information by the DFG (German Research Foundation), GRK 1498.

REFERENCES

ANGELSEN, A. Deforestation: population or market driven? **Development Studies and Human Rights**, Working Paper, p.42. 1997.

BECKER, B. Geopolítica da Amazônia. **Journal of the Institute of Advanced Studies of the University of Sao Paulo**, vol.19, pp.71–86. 2005.

BRITANNICA. **Britannica Online Encyclopedia**. Available online at: http://www.britannica.com/. 2008

HANSEN, M., Y. SHIMABUKURO, P. POTAPOV and K. PITTMAN. Comparing annual MODIS and PRODES forest cover change data for advancing monitoring of Brazilian forest cover. **Remote Sensing of Environment**, vol. 112,no. 10, October 2008, pp. 3784–3793. 2008.

INPE. PRODES - Amazon deforestation database: São Jose dos Campos, INPE. Available online at: www.obt.inpe.br/prodes. 2009.

KUHN, W. Geospatial semantics: Why, of what, and how? **Journal on Data Semantics III**, pp. 1–24. 2005.

MACHADO, L. O. A fronteira agrícola na Amazônia brasileira. In: A. CHRISTOFOLETTI, B. K. BECKER*, et al* (Ed.). **Geografia e meio ambiente no Brasil**. São Paulo: Companhia das Letras, 1998. A fronteira agrícola na Amazônia brasileira, pp.181–217.

SHIMABUKURO, Y., V. DUARTE, L. ANDERSON, D. VALERIANO, E. ARAI, F. R., B. RUDORFF and M. MOREIRA. Near real time detection of deforestation in the BrazilianAmazon using MODIS imagery. **Ambiente & Água- An Interdisciplinary Journal of Applied Science**. 2007.