Undefined 0 (0) 1
IOS Press

Linked Brazilian Amazon Rainforest Data

Tomi Kauppinen [a], Giovana Mira de Espindola [b], Jim Jones [a], Alber Sánchez [b], Benedikt Gräler [a], Thomas Bartoschek [a]

[a] *Institute for Geoinformatics, University of Muenster, Germany*
[b] *Earth System Science Center, National Institute for Space Research (INPE), Brazil*

Abstract. The Linked Brazilian Amazon Rainforest Data contains observations about deforestation of rainforests and related things such as rivers, road networks, population, amount of cattle, and market prices of agricultural products. The Linked Data approach offers thus to combine ecological, economical and social dimensions together. Our aim has been to 1) dramatically shorten the time needed to collect information for a research setting concerning the Brazilian Amazon, and 2) via the linkage between datasets enable novel types of transdisciplinary research for the scientific community.

Keywords: Brazilian Amazon, Linked Open Data, Deforestation, Spatio-temporal Datasets

1. Introduction

Open Science needs Open Data to maximize the transparency, reproducibility and reuse of scientific efforts. An example of a high demand for data is the research about climate change, for example about the role of deforestation in it.

Deforestation and its related phenomena such as market prices of agricultural products form together a complex system. There is an urgent need to share and publish research data about it, as it would enable other researchers to interconnect their data to the published ones. The benefit is that these explicit interconnections allow for the analysis of all of the resulting linked data in a transdisciplinary manner. Thus the whole complex socio-economic and environmental system could be modelled and not just subsets of it.

In this paper our contribution is to describe how large amounts of remote sensing observation data about the Brazilian Amazon Rainforest has been published as Linked Spatiotemporal Data. The data covers the whole Brazilian Amazon Rainforest. Moreover, we show how this data can be further accessed and analyzed using R statistical computing environment by openly available methods via a tutorial. We argue that this is a contribution towards Linked Science[3,2], where not just publications, but data, methods, tools, and other scientific assets are interconnected and shared online. The work is a continuation of our earlier work[4], but with substantial additions to the published dataset.

2. Linked Brazilian Amazon Rainforest Data

2.1. Linking a Diverse Variety of Data Together

Governments maintain a diverse amount of different registers—e.g. about population, export, employment, river structures, etc.—for decision making. These data are also very valuable in statistical analysis for finding correlations between different phenomena. Open and broad access in a uniform way to such data eases and enables scientific research.

Our aim is to create a unique dataset which we call the Linked Brazilian Amazon Rainforest Data (LBARD). The goal is to enhance research about deforestation. Motivation is that while the Brazilian government and increasingly also other authorities provides public access to data as spreadsheets, the linkage between these data is missing.

Thus this creates challenges to conduct e.g. time-series or spatial analysis.

2.2. Study Area

The study area is the Brazilian Amazon rainforest, which covers an area of more than 5 million square kilometers. Data representing deforestation, land uses— pasture, temporary and permanent agriculture—and natural and social factors of change were aggregated to grid cells of 25 km x 25 km, counting a total of 8580 cells. For each cell of the resulting grid, there are 38 natural and social factors available, grouped into eight categories—land use, demography, environmental, accessibility to markets, technology, public policy, market pressure and agrarian structure. Each of the variable was described similarly as below the variable DEFOR_2004 (which is used to describe the percentage of new deforestation in 2004 for each grid cell).

```
@prefix amazon:
   <http://spatial.linkedscience.org/context/amazon/> .

amazon:DEFOR_2004
      rdf:type amazon:VARIABLE ;
      rdfs:label "Percentage of new deforestation in 2004"^^xsd:string ;
      amazon:aggregation amazon:Pixel ;
      amazon:timeperiod amazon:year2004 ;
            rdf:type amazon:TIMEPERIOD ;
                lsv:begin "2007-01-01" ;
                lsv:end   "2007-12-31" .
      amazon:unit amazon:percent ;
      amazon:variabletype amazon:LandUse .
      amazon:source amazon:INPE ;
            rdf:type amazon:SOURCE ;
                foaf:homepage
                    <http://www.inpe.br> .
```

2.3. Describing Deforestation and Land Uses

We made use of the Landsat TM-based 1997-2007 deforestation maps produced under the Amazon-monitoring program of the Brazilian National Institute for Space Research (INPE) in year 2010. The percentages of accumulated deforestation in different years were computed for each grid cell. The accumulated deforestation in 1997 and 2007 was decomposed into the main agricultural uses—pasture, temporary and permanent agricultures. Economical data provided by Brazil's Informa Economics FNP, number of conservation units provided by the Brazilian Ministry of Environment (MMA), and distances to road structures from data by the Instituto Brasileiro de Geografia e Estatística (IBGE) were aggregated to the grid level in a similar manner.

The excerpt below shows an example grid cell with accumulated and yearly deforestation values.

```
@prefix amazon:
   <http://spatial.linkedscience.org/context/amazon/> .

amazon:AMZ_LINKED_25K_1000
      rdfs:label "Cell 1000"^^xsd:string ;
      amazon:ACUM_1997 "0.496"^^xsd:double ;
      amazon:ACUM_2002 "0.61"^^xsd:double ;
      amazon:ACUM_2007 "0.703"^^xsd:double ;
      amazon:ACUM_2008 "0.706"^^xsd:double ;
      amazon:DEFOR_2002 "0.039"^^xsd:double ;
      amazon:DEFOR_2003 "0.0030"^^xsd:double ;
      amazon:DEFOR_2004 "0.031"^^xsd:double ;
      amazon:DEFOR_2005 "0.042"^^xsd:double ;
      amazon:DEFOR_2006 "0.0050"^^xsd:double ;
      amazon:DEFOR_2007 "0.012"^^xsd:double ;
      amazon:DEFOR_2008 "0.0040"^^xsd:double ;
      ...
```

2.4. Processing and Describing Data from Data Portals

In order to process and describe other statistical data[1] such as census data, we developed an application capable of downloading, and transforming spreadsheets into Linked Data. The process made use of the Open Linked Amazon (OLA) vocabulary tailored for this purpose, and a number of established vocabularies. Open Linked Amazon vocabulary assembles classes and properties describing specific measuring units and variables used for the Brazilian government for publishing records. The time series are built on observations of a specific variable in a certain time period. Below is an example excerpt from the data.

```
@prefix amazon:
   <http://spatial.linkedscience.org/context/amazon/> .

amazon:BRAZIL_MUNICIPALITY_110092
   amazon:hasObservation
      amazon:OBS_CATTLE_2007_1 .
amazon:OBS_CATTLE_2007_1
   amazon:isAbout
      amazon:Cattle .
   amazon:year
      "2007"^^xsd:integer .
   amazon:amountProduced
      "297586"^^xsd:integer .
   amazon:hasUnit
      amazon:Head .
```

In the sample below an observation of Temporary Crop of Soy Beans in the year of 2004. This observation represents the amount of soy beans production in metric tons regarding to the municipality of Uruará within the Brazilian Amazon.

```
@prefix amazon:
   <http://spatial.linkedscience.org/context/amazon/> .
@prefix xsd:
```

[1] As collected by Instituto Brasileiro de Geografia e Estatística (IBGE)

```
   <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf:
   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dbpedia:
   <http://dbpedia.org/resource/> .


amazon:BRAZIL_MUNICIPALITY_150815
    amazon:hasObservation
        amazon:OBS_TEMP_CROP_2004_SOYBEAN_691 .
amazon:OBS_TEMP_CROP_2004_SOYBEAN_691
        amazon:isAbout dbpedia:Soybean .
amazon:OBS_TEMP_CROP_2004_SOYBEAN_691
        amazon:year "2004"^^xsd:integer .
amazon:OBS_TEMP_CROP_2004_SOYBEAN_691
        rdf:type amazon:TemporaryCrop .
amazon:OBS_TEMP_CROP_2004_SOYBEAN_691
        amazon:amountProduced "840"^^xsd:integer .
amazon:OBS_TEMP_CROP_2004_SOYBEAN_691
        amazon:hasUnit dbpedia:Ton .
```

2.5. Enriching Data with Spatial Relationships

The spatial aspects of data were Modeled using the Open Time and Space Core Vocabulary (TISC)[2]. Missing classes and properties were added as an extension of TISC. For example, in order to enabling spatial data aggregation capabilities, a structure enabling to express the amount of partial overlap between different regions.

This is critical for relating a study area grid and other information available with information—e.g. the population of a municipality—from a different aggregation level. This modeling decision allows for estimating the value of a variable when aggregated spatially between different granularities. An example is a grid's cell overlapping two or more different municipalities. If an approximation of the population density is needed at the grid level, then properties like the overlap ratio between a grid and a municipality enable to aggregate the population data for approximations at the grid level. Below is an excerpt of this data to give an example.

```
amazon:PARTIALLY_OVERLAY_FROM_AL25K1027_TO_BRMUN5218003
    tisc:partialOverlapFrom
        amazon:AMZ_LINKED_25K_1027 ;
    tisc:partialOverlapTo
        r-town:BRAZIL_MUNICIPALITY_5218003 ;
    tisc:partialOverlapArea
        "477.22"^^xsd:double ;
    tisc:partialOverlapUnit
        dbpedia:Square_kilometre ;
    tisc:partialOverlapRatio
        "0.76"^^xsd:double .
```

As a result of the dataset contains the following data:

Amount of Deforestation The observed deforestation aggregated to grid cells (see [1] for details of how the aggregation was done).

Cattle The total number of heads of cattle grouped by municipality.

Legal Amazon Grid This grid covers the whole Legal Amazon area divided in 8480 cells with the size of 25km x 25km.

Municipalities and Federal States A complete compound of Brazilian Federal States which belong to the Legal Amazon area and their municipalities, together with their geographical location and covered area.

Permanent Crops Crops of produced from specific plants which last for many seasons classified by hectares of planted and harvested area, thousand reais, metric tons and kg/hectare. The permanent crops are linked to an observation and additionally its plants are linked to their respectively entries in DBPedia[3].

Population per municipality Census 2000 and 2010, and population projections from 2001 to 2009.

Temporary Crops Crops of produced from specific plants which last for less than one year classified by hectares of planted and harvested area, thousand reais, thousand fruits, metric tons and kg/hectare. The temporary crops are linked to an observation and additionally its plants are also linked to their respectively entries in DBPedia.

Rivers Rivers which geographically overlap a Legal Amazon Grid cell.

River Basins River basins which geographically overlap a Legal Amazon Grid cell.

The basic information of the dataset is as follows:

- Name: Linked Brazilian Amazon Rainforest Data (LBARD)
- Version date: 2012-05-16
- Version number: 2.0
- Licensing: Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
- Availability: As Linked Data on the web: http://linkedscience.org/data/linked-brazilian-amazon-rainforest/

2.6. Metrics and statistics

The table 1 reports the variety and amounts of data that was linked together, to external datasets, and which vocabularies were in use for different types of data.

[2] http://observedchange.com/tisc/ns/

[3] http://dbpedia.org

	Total Triples	Total ext. Links to DBPedia	Vocabs in Use
Temporary crops	1420755	608895	OLA
Permanent crops	1420755	608895	OLA
Grid cells	8580	0	OLA
Municipality coordinates	5799	0	TISC
Municipality type	5799	5799	TISC
River definition	170	0	RDF
River coordinates	85	0	RDF
River type	85	0	RDF
Overlay cells/municipalities	89115	17823	TISC
Overlay cells/rivers	8580	0	TISC
Mesoregions	18	6	RDF
Microregions	88	33	RDF
Municipality code mappings	5565	0	OWL
Municipality/state partonomy relations	16797	0	TISC
Overlap municipality vs. rivers	36165	7233	TISC
Partial overlap cells vs. municipalities	17823	0	TISC
Partial overlap instances	17823	0	TISC
Municipality/microregion relations	143	0	TISC
Municipality to DBPedia relations	136	136	OWL
Total	3338432	1370588	

Table 1

Metrics of the variety and amounts of data that was linked together and to external datasets.

3. Example Use Case: Accessing the Data from R

One crucial aspect is how to access and analyse data, and especially how to get only that part of data which is of interest for a given research question. Linked Data solves the access part, and SPARQL allows to query only a subset of the data. For statistical computing there are tools like R[4], and a separate package[5] [5] for it supports querying Linked Data.

As an example use case we provide an online tutorial[6] to explore the data from R and plot it on maps. The aim here is support bridging of the two communities, those of statistical computing and the semantic web.

[4] http://www.R-project.org/

[5] http://linkedscience.org/tools/sparql-package-for-r/

[6] see http://linkedscience.org/tutorials/

4. Shortcomings and Future Work

One of the original plans was to use the The Statistical Core Vocabulary (SCOVO)[7] for describing the data, but this turned out to be challenging for various reasons. The main problem was the However, we are currently exploring the possibility of using RDF Data Cube vocabulary[8], and plan to describe the data using it, in addition to the already published data. Thus we acknowledge that the lack of using RDF Data Cube vocabulary is a shortcoming of our current approach as it restricts the linkage via the shared models to other datasets.

However, we showed in this paper that we provide 1) an extensive amount of external linkage, and also 2) make use common vocabularies, and that 3) the published dataset as such can be used for extensive analysis of related phenomena concerning the Brazilian Amazon Rainforest. In the future work we plan to extend the dataset further

[7] http://vocab.deri.ie/scovo

[8] http://www.w3.org/TR/vocab-data-cube/

by including diverse open data concerning ecological, economical and social aspects related to rainforests. Moreover, we are currently developing a set of applications that make use of the data. We hope to learn via these experiments and evaluations about both the potential and the future development needs of the dataset.

5. Conclusions

In this paper we described the Brazilian Linked Rainforest Data. We argue that the version 2.0 of the Brazilian Amazon Linked Rainforest Data to be major step in having open and linked data about the ecological, economical and social dimensions related to the Brazilian Amazon Rain forest. The data has been published using the Linked Data principles. The emerging application scenarios such as statistical analysis of the data show that the dataset has a practical value.

Acknowledgements

This research has been partially funded by the International Research Training Group *Semantic Integration of Geospatial Information* (DFG GRK 1498).

References

[1] Giovana Mira de Espindola. *Spatiotemporal trends of land use change in the Brazilian Amazon*. PhD thesis, National Institute for Space Research (INPE), São José dos Campos, Brazil, 2012.

[2] Tomi Kauppinen, Alkyoni Baglatzi, and Carsten Keßler. Linked Science: Interconnecting Scientific Assets. In Terence Critchlow and Kerstin Kleese-Van Dam, editors, *Data Intensive Science*. CRC Press, USA, forthcoming 2012.

[3] Tomi Kauppinen and Giovana Mira de Espindola. Linked Open Science—communicating, sharing and evaluating data, methods and results for executable papers. *Proceedings of the International Conference on Computational Science (ICCS 2011), Procedia Computer Science*, 4(0):726–731, 2011.

[4] Tomi Kauppinen, Giovana Mira de Espindola, and Benedikt Gräler. Sharing and analyzing remote sensing observation data for Linked Science. In *Poster proceedings of the Extended Semantic Web Conference 2012 (ESWC2012)*, Heraklion, Crete, Greece, May 2012.

[5] Willem Robert van Hage and Tomi Kauppinen. SPARQL Package for R, May 22 2012.