DBCells – an open and global multi-scale linked cells

Sérgio Souza Costa[1], **Evaldinolia Gilbertoni Moreira**[2],
Micael Lopes da Silva[1], **Thamyla Maria de Sousa Lima**[1]

[1]Curso de Engenharia da Computação – Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brazil

[2]Departamento Acadêmico de Informática – Instituto Federal do Maranhão (IFMA)
São Luís – MA – Brazil

sergio.costa@ufma.br, evaldinolia@ifma.edu.br

micaelopes32@gmail.com, thamyla.sl@gmail.com

Abstract. *The land change models require large amounts of data, are difficult to be reproduced, as well as to be reused. Some initiatives to open and link data increase the reproducibility of scientific experiments and data reuse. One pillar of the linked data concept is the use of Uniform Resource Identifier (URI). In this paper, we propose DBCells – an architecture for publication of a global cellular space where each cell has a URI. This new approach will allow comparison, reproduction and the reuse of models and data. However, in order to succeed, this proposal requires participation, partnerships and investments. Our main purpose in this paper is to present the architecture, benefits and challenges for debating with the scientific community.*

1. Introduction

The reproducibility is a crucial characteristic for experimental science and requires access to data and tools [Molloy 2011]. Furthermore, the comparison and reuse of data and results play critical roles. The achievement of these requirements is a great challenge in experiments that demand large volumes of data, like land change models. These models demand data from environmental, social, technological, and political drivers [Moran et al. 2005, Turner et al. 2007]. In general, each driver is represented as a value into spatial unit, pixel or cell. A pixel is the smallest addressable element in the raster layer that represents a spatial variable, like slope or distance to roads. The cell space is an alternative representation, where each cell handles one or more types of attribute [Câmara et al. 2008]. In both cases, cells and pixels are not treated as unique and distinct entities, but as partitions of a continuous space. Then, even the smallest differences in the bounding box of the study area can generate different cell spaces. These differences make the comparison and reuse of data a great challenge. In this paper, we propose that each cell from each resolution is a unique and distinct entity that has a universal identifier, what we call DBCells architecture.

The Uniform Resource Identifier (URI) is one of the pillars of the web data architecture, which links data instead of pages. The architecture proposed by Tim Berners-Lee is referred to as linked data [Berners-Lee 2006] and provides support for large datasets, such as DBpedia [Auer et al. 2007] and GeoNames [Wick and Vatant 2012]. The DBpedia describes all the concepts from wikipedia through URI, for example, the National Institute for Space Research is identifiable by `http://dbpedia.org/data/`

`National_Institute_for_Space_Research.rdf`. This institute is located in São José Dos Campos, and is identified in the GeoNames by the following url: `http://sws.geonames.org/6322578/about.rdf`.

Several authors have argued that linked data allows experiments to become more reproducible, which depends on large volumes of data [Kauppinen and De Espindola 2011, Molloy 2011]. In addition, some authors argue that linked data can be explored to share large volumes of data among the scientific community [Quoca et al. 2014, Baučić and Medak 2014]. In [Quoca et al. 2014] the authors describe how NOAA dataset can be transformed and published as linked data. The data from 20.000 weather sensor stations over the world were converted to 177 billion triples. Other example of linked open data is the Linked Brazilian Amazon Rainforest Data [Kauppinen et al. 2014]. This dataset is openly available for anyone as non-commercial research use. However, in this dataset each variable (land use, demography, environmental, accessibility to markets technology) is strongly coupled to the cells. In our architecture proposal, described in Section 3, the cells are distinct entities that have an universal identifier, which can be linked from other data. In other words, each cell is a spatial unit that can link results and data from land change models. This paper is organized as follows: Section 2 presents the two major concepts – the open linked data and cellular-space; Section 3 describes DBCells – the architecture proposed; Section 4 summarizes the main benefits and challenges to achieve the link between the models in global scale.

2. Theoretical foundation

2.1. Open linked data

First of all, it is necessary to distinguish data, linked data and open data, shown in Figure 1. Data are the base of the pyramid, and are defined as symbols that represent properties of objects, events and their environment [Ackoff 1989]. Open data are all those that can be freely used, modified, and shared by anyone for any purpose [The Open Definition 2013]. The linked data refers to a set of best practices for publishing and interlinking structured data on the Web [Heath and Bizer 2011].

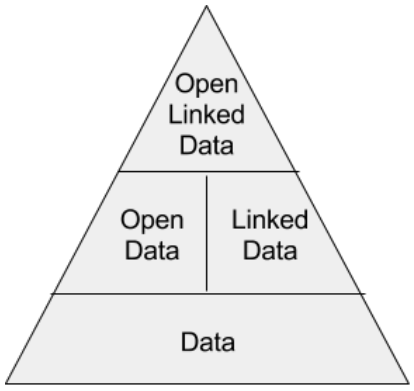

Figure 1. From data to open linked data

The movement of open data is inspired by the open source and consists of three

major concepts: openness, participation and collaboration [Chignard 2013]. These concepts are present in the following three key features: (a) Availability and access – the data must be available as a whole and in a way that does not create complicated processes for the interested party in copying it; (b) Reuse and Redistribution – the data must be provided under terms that permit reuse and redistribution, including combining this data with other datasets; (c) Universal Participation – everyone must be able to use, reuse and redistribute; there should be no discrimination against fields of endeavour or against persons or groups [Dietrich et al. 2009]. The open data movement can bring democratic gain, like better transparency of public action, citizen participation and response to the crisis of confidence towards politicians and institutions [Chignard 2013, Janssen et al. 2012]. However, authors point out some prerequisites: the availability on the web and the machine readable. In other words, they must follow the three laws proposed in [Eaves 2009], which are:

1. If the data cannot be spidered or indexed, it does not exist;
2. If the data is not available in open and machine readable format, it cannot engage;
3. If a legal framework does not allow it to be repurposed, it does not empower.

Being readable for the machines is also one of the characteristics required for the linked data. However, in the linked data concept, it is necessary to link and allow it to be linked by other datasets, which is summarized in the following principles [Berners-Lee 2006]:

1. Use URIs as names for things;
2. Use HTTP URIs, so that people can look up those names;
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL);
4. Include links to other URIs, so that they can discover more things.

In [Berners-Lee 2006], the author describes the datasets in terms of five-stars. Each rating represents a progressive transition from data to Linked Data. Every data is available on the web (at any format), but the ones with an open license have 1 star. In addition to that, if the data is available as machine-readable structured data (e.g., Microsoft Excel instead of a scanned image of a table) then it has 2 stars. To have 3 stars, the data needs to be available at a non-proprietary format (e.g., CSV instead of Excel). The next star requires the data to be available according to the previous constraints, plus the use of open standards from the W3C (RDF and SPARQL), in order to identify things, so that people can link to it. Finally, to have 5 stars, data needs to be available according to all the above criteria, plus to provide context via outgoing links to other people's data. It is important to emphasize that the opening is not a prerequisite for linked data. For example, a private company can link their data, but does not necessarily make them open. Figure 2 shows the linking open data cloud in 2014. The DBpedia [Auer et al. 2007] and GeoNames [Wick and Vatant 2012] datasets are located in the center.

The open and linked data is an important element to open science [Murray-Rust 2013, Kauppinen and De Espindola 2011]. In [Kauppinen and De Espindola 2011], the authors propose the Linked Open Science aiming to be a standardized and generic recipe for executable papers. This concept was built on these four key elements: (a) Linked Data, (b) OpenSource and Web-based Environments, (c) Cloud Computing and (d) Creative Commons. An example of linked

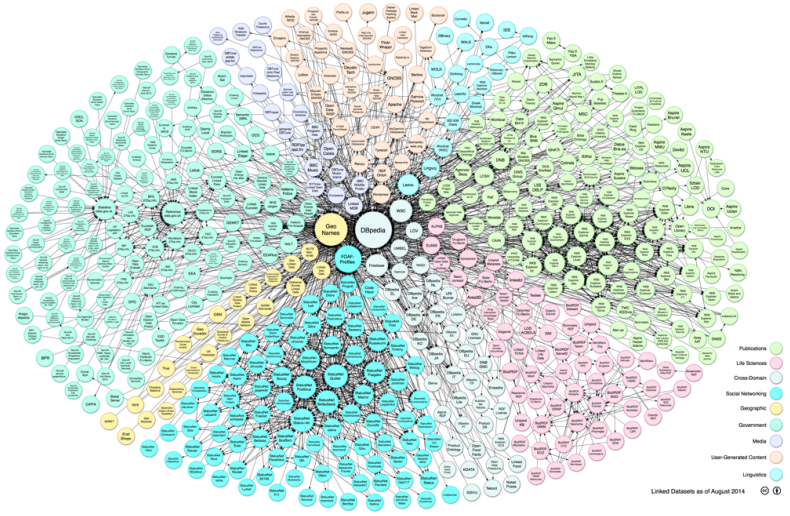

Figure 2. Linking open data cloud. Source: [Cyganiak and Jentzsch 2014]

open data is the Linked Brazilian Amazon Rainforest Data [Kauppinen et al. 2014]. This dataset is openly available for anyone that will make non-commercial research use of it. The data was produced by the Institute for Geoinformatics, University of Muenster, Germany and the National Institute for Space Research (INPE) in Brazil. However, in this dataset, each variable (land use, demography, environmental, accessibility to markets technology) is strongly coupled to the cells. In our proposal, described in the Section 3, the cells are distinct entities that have a universal identifier, which can be linked from other datasets.

2.2. From geo-fields to cellular space

Our focus is on data from land change models. In general, these models describe phenomena that vary continuously in space and time, as deforestation in the Brazilian Amazon region. Their input and output are represented as geo-fields. Together with geo-objects, the geo-fields are the two fundamental spatial representations [Kuhn 2012, Câmara 2005]. Geo-objects describe entities that have an identity as well as spatial, temporal, and thematic properties [Kuhn 2012]. However, geo-fields have been shown to be more fundamental than geo-objects and are capable of integrating both representations [Liu et al. 2008, Camara et al. 2014, Costa et al. 2007]. As data structure, the geo-fields are discretized and used two ways [Kuhn 2012]:

1. through a finite number of cells, within each one the attribute is assumed to remain constant;
2. through a finite set of sample points with interpolation rules for positions among them.

In this paper, we are interested in the first way, where the study area is partitioned forming a regular grid of square, triangular, hexagonal, or cubic cells as in raster based layers or a cellular space. The raster model can be compared with a bitmap image, which consists of a number of pixels organized in rows and columns. Basically, in most cases, raster data is indeed derived from satellite images, which serve as a basis for observing weather, vegetation or electromagnetic radiation. A cellular space is an alternative

model to represent geo-fields. It is a spatial data type where each cell handles one or more types of attribute [Câmara et al. 2008]. In [Câmara et al. 2008],the authors argue that cellular spaces were part of the early GIS implementations, but now it is time to reconsider this decision and reintroduce it as a basic data type; they also argue that the usage of one-attribute raster data in the storage of results for dynamical models requires the storage of information in different files. By the other hand, a cellular space stores all attributes of a cell together, with significant benefits for modeling, in contrast to the more cumbersome single value raster approach. Together with the concept of Generalized Proximity Matrix (GPM), it is possible to represent hierarchical and network relations [de Aguiar et al. 2003] and [Moreira et al. 2008]. In [Moreira et al. 2008], the authors use these concepts to represent hierarchical and network spatial relations in multi-scale land change models.

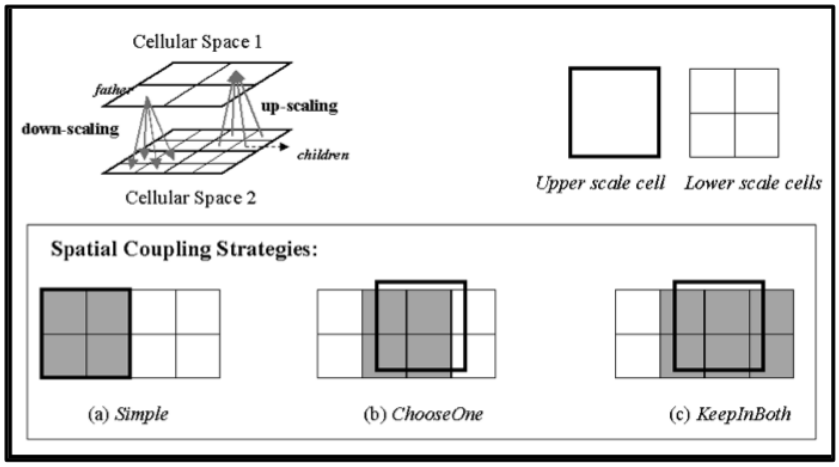

Figure 3. Representation of strategies for spatial coupling in the case of regular cells. Source: [Moreira et al. 2008]

Cellular spaces have been used for simulation of urban and environmental models as part of cellular automata models (Batty 2000). In TerraLib ([Câmara et al. 2008]) and TerraME ([de Senna Carneiro et al. 2013]) the cellular space is a native building block. These concepts and tools have supported the development of models published in the literature [Aguiar et al. 2007, Moreira et al. 2009, Aguiar et al. 2012, Espindola et al. 2012, Andrade et al. 2009].

3. The architecture

We propose a layered architecture (in five layers), adapted from [Heath and Bizer 2011], as showed in Figure 4. The publication layer includes a dataset of cellular space and multi-scale relationships. The web of data layer links the cellular space to existing datasets, like Geonames, DBPedia and SWEET Ontology. The data access and storage layer integrate local and web data, providing a transparent access and storage for modeling tools. The model layer uses and shares data provided by the lower layers. For example, coarser scales run models of climate, and finer scales run environmental and social models. The user layer runs and reproduces the experiment of a particular model. A user can pub-

lish the results and the data of an experiment in the web of data, allowing replication of experiments, an essential characteristic of science.

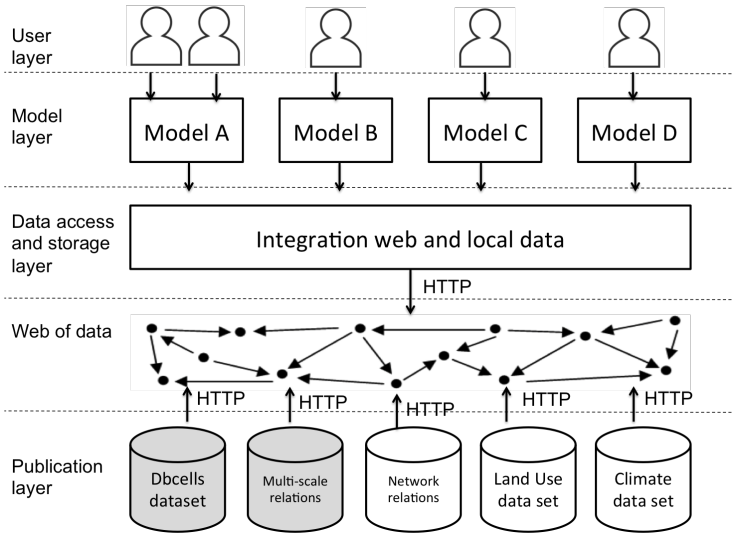

Figure 4. The DBCells Architecture Proposal

In the dataset of cellular space, each cell is identified by a URI and described as a RDF graph, see Figure 5. The RDF (Resource Description Framework) is the data model, standardized by W3C for representing Semantic Web resources. It expresses information as graphs consisting of triples with subject, property and object [Klyne and Carroll 2006]. These three graph elements are identifiable through URI. In the dataset of cellular space, each graph consists of minimal set of properties to describe a cell, such as its position and bounding box. These graphs can be stored in a graph database, like the Neo4J[1], and serialized as RDF/XML, see Code 1.

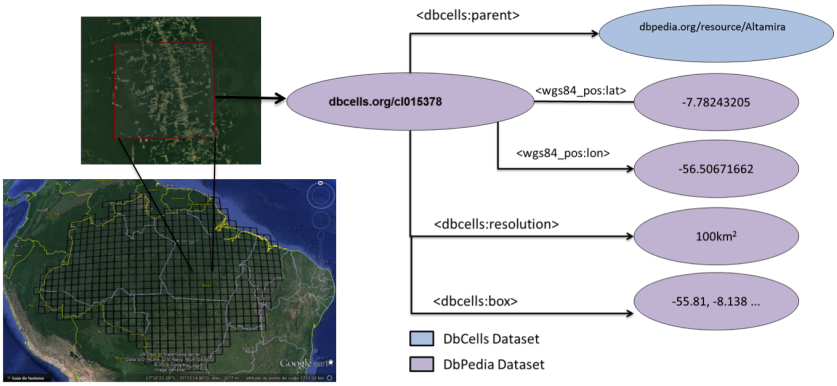

Figure 5. A specific cell as a RDF graph

[1] https://neo4j.com/

```
1
2  ...
3  <rdf:Description rdf:about="http://dbcells.org/cl015378">
4          <wgs84_pos:lat rdf:datatype="http://www.w3.org/2001/
               XMLSchema#double">-7.782432055255575</wgs84_pos:lat>
5          <wgs84_pos:long rdf:datatype="http://www.w3.org/2001/
               XMLSchema#double">-56.50671662245421</wgs84_pos:long
               >
6          <parent xmlns="http://dbcells.org/">
7                  http://dbpedia.org/resource/Altamira</parent>
8          <resolution xmlns="http://dbcells.org/">100000</
               resolution>
9          <dbcells:box>
10                 -55.81526040160506,-8.138132254243271,0
                       -54.90595645484754,-8.138133321360099,0
                       -54.90595545105769,-7.2313407607591,0
                       -55.81525938511868,-7.23133981183386,0
                       -55.81526040160506,-8.138132254243271,0
11         </dbcells:box>
12 </rdf:Description>
13 ...
```

Code 1. A specific cell graph serialized as a RDF/XML

Our proposal is to describe both, the cells and their relationships, through RDF graphs. Graphs express different relations, including: (a) topological relations; (b) network connectivity, both physical (e.g., transportation infrastructure) and logical (e.g., trade fluxes); (c) vicinity in cell spaces and grids; (d) coupling between spatial scales [Moreira et al. 2008]. We propose to describe the relationships in different datasets, allowing a model to select one or more relationships. The Figure 6 shows an example where the relationships describes a spatial coupling between cellular spaces in different resolutions. In this case, the nodes are cells and the edges represent the hierarchical relations. Similarly, these relationships can be stored in the graph database and serialized as RDF / XML.

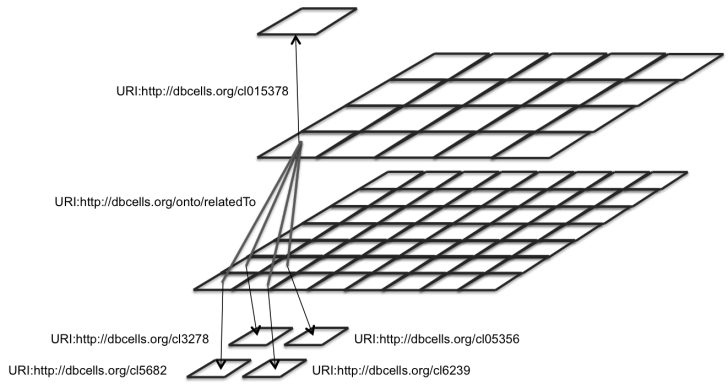

Figure 6. Relationships between multi-scale cellular spaces as a graph

The DBCells architecture is under development, and will require partnerships and investments. This article is intended to present and validate this proposal together with the scientific community, as its success will depend on the interest of this community. In

the next section we present some benefits and challenges to complete this project.

4. Benefits and challenges

The implementation of this proposal brings several benefits and challenges. Similar to DBPedia and GeoNames, the DBCells may be a dataset that will link datasets from different spatial models. Since each cell has a universal identifier, the models can link their data and results to it. This will allow sharing data and results, and the reuse of datasets already published, illustrated in Figure 7.

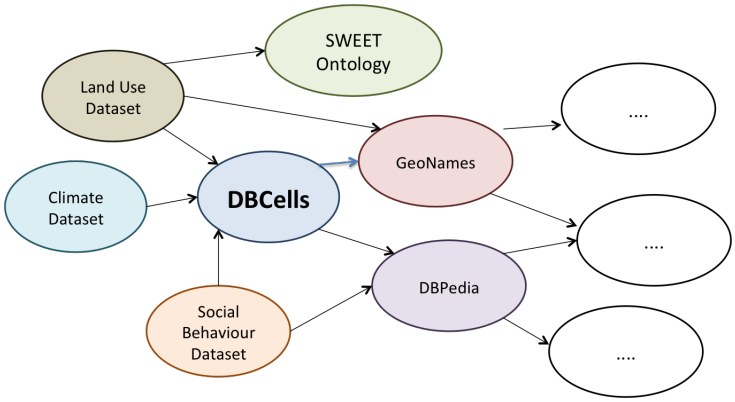

Figure 7. Integration between datasets

The open data is crucial for reproducibility of data demanding experiments [Murray-Rust 2013, Kauppinen and De Espindola 2011, Molloy 2011].

According to [Molloy 2011]:

> *"The more data is made openly available in a useful manner, the greater the level of transparency and reproducibility and hence the more efficient the scientific process becomes, to the benefit of society".*

The relationships between cellular spaces allow the reuse of models at different scales and resolutions. For example, a land use model at a finer scale can use results of a climate model in a coarser scale, which is represented in Figure 8.

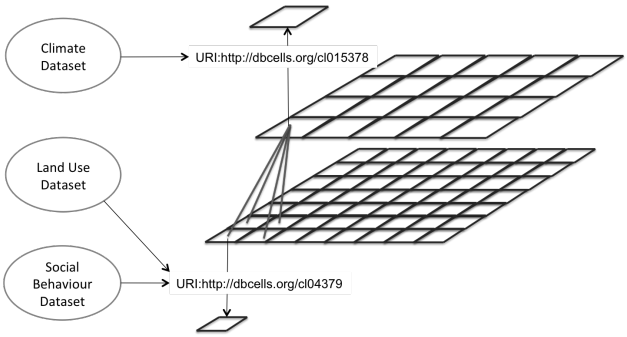

Figure 8. Reuse of model data and results between cellular spaces

The proposed architecture will also contribute to a better reproducibility and comparison of the data demanding experiments. This benefit is enhanced through open source environment tools [Kauppinen and De Espindola 2011]. This architecture presents also several challenges, for example, the participation of the scientific community. The benefits previously mentioned will depend on the community interest in making their data open and linked. Furthermore, each modeling tool will need to implement the data access and storage layer to retrieve and store data on the web. Another challenge is the conflict between vocabularies from different models. Therefore, it will be necessary to use the already established vocabularies, whenever possible. At last, an efficient and distributed computing will be necessary for storage and retrieval of data from a global cellular space at different scales. For that reason, we will conduct the initial experiments in areas of greatest interest by the scientific community like Amazon rainforest.

5. Final remarks

This paper introduced an innovative architecture – DBCells – that integrates two concepts: cellular spaces and linked data. The pillar of integration is to treat each cell as a unique and distinct entity that has a universal identifier. To achieve this integration, we propose four steps: 1) divide the space in regular cells, 2) associate each cell to an identifier, 3) represent each cell as an RDF graph available on the web and 4) connect data and results models to these identifiers. The main benefits of the new approach are the reuse, sharing, comparison and reproduction of land change models. The main challenges are the participation and interest of the scientific community, and an efficient architecture to store and retrieve large volume of data. Thus, the success of this proposal requires partnerships and investments. Based on that, by presenting our vision, we expect to raise an engaging debate with the scientific community.

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1).

Aguiar, A. P. D., Câmara, G., and Escada, M. I. S. (2007). Spatial statistical analysis of land-use determinants in the brazilian amazonia: Exploring intra-regional heterogeneity. *Ecological modelling*, 209(2):169–188.

Aguiar, A. P. D., Ometto, J. P., Nobre, C., Lapola, D. M., Almeida, C., Vieira, I. C., Soares, J. V., Alvala, R., Saatchi, S., Valeriano, D., et al. (2012). Modeling the spatial and temporal heterogeneity of deforestation-driven carbon emissions: the inpe-em framework applied to the brazilian amazon. *Global Change Biology*, 18(11):3346–3366.

Andrade, P. R., Monteiro, A. M. V., Câmara, G., and Sandri, S. (2009). Games on cellular spaces: How mobility affects equilibrium. *Journal of Artificial Societies and Social Simulation*, 12(1):5.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Baučić, M. and Medak, D. (2014). Building the semantic web for earth observations. In *DailyMeteo. org/2014 Conference*.

Berners-Lee, T. (2006). Linked data.

Camara, G., Egenhofer, M. J., Ferreira, K., Andrade, P., Queiroz, G., Sanchez, A., Jones, J., and Vinhas, L. (2014). Fields as a Generic Data Type for Big Spatial Data. *Geographic Information Science*, page in press.

Câmara, G., Vinhas, L., Ferreira, K. R., De Queiroz, G. R., De Souza, R. C. M., Monteiro, A. M. V., De Carvalho, M. T., Casanova, M. A., and De Freitas, U. M. (2008). Terralib: An open source gis library for large-scale environmental and socio-economic applications. In *Open source approaches in spatial data handling*, pages 247–270. Springer.

Chignard, S. (2013). A brief history of open data. *Paris Tech Review*, 29.

Costa, S. S., Câmara, G., and Palomo, D. (2007). Terrahs: integration of functional programming and spatial databases for gis application development. In *Advances in Geoinformatics*, pages 127–149. Springer.

Cyganiak, R. and Jentzsch, A. (2014). Linking open data cloud diagram. *LOD Community (http://lod-cloud. net/)*, 12.

Câmara, G. (2005). Representação computacional de dados geográficos. *CASANOVA, MA et al. Banco de dados geográficos. Curitiba: Mundogeo*, pages 11–52.

de Aguiar, A. P. D., Câmara, G., Monteiro, A. M. V., and de Souza, R. C. M. (2003). Modelling spatial relations by generalized proximity matrices. In *GeoInfo*.

de Senna Carneiro, T. G., de Andrade, P. R., Câmara, G., Monteiro, A. M. V., and Pereira, R. R. (2013). An extensible toolbox for modeling nature–society interactions. *Environmental Modelling & Software*, 46:104–117.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, P., Tait, J., and Zijlstra, T. (2009). Open data handbook.

Eaves, D. (2009). The three laws of open government data.

Espindola, G. M., De Aguiar, A. P. D., Pebesma, E., Câmara, G., and Fonseca, L. (2012). Agricultural land use dynamics in the brazilian amazon based on remote sensing and census data. *Applied Geography*, 32(2):240–252.

Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4):258–268.

Kauppinen, T. and De Espindola, G. M. (2011). Linked open science-communicating, sharing and evaluating data, methods and results for executable papers. *Procedia Computer Science*, 4:726–731.

Kauppinen, T., De Espindola, G. M., Jones, J., S??nchez, A., Gr??ler, B., and Bartoschek, T. (2014). Linked Brazilian Amazon Rainforest Data. *Semantic Web*, 5(2):151–155.

Klyne, G. and Carroll, J. J. (2006). Resource description framework (rdf): Concepts and abstract syntax.

Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276.

Liu, Y., Goodchild, M. F., Guo, Q., Tian, Y., and Wu, L. (2008). Towards a general field model and its order in gis. *International Journal of Geographical Information Science*, 22(6):623–643.

Molloy, J. C. (2011). The open knowledge foundation: open data means better science. *PLoS Biol*, 9(12):e1001195.

Moran, E., Ojima, D., Buchmann, N., et al. (2005). Global land project: Science plan and implementation strategy. *IGBP Report*, 33.

Moreira, E., Costa, S., Aguiar, A. P., Câmara, G., and Carneiro, T. (2009). Dynamical coupling of multiscale land change models. *Landscape Ecology*, 24(9):1183–1194.

Moreira, E., de Aguiar, A. P. D., Costa, S. S., and Câmara, G. (2008). Spatial relations across scales in land change models. In *GeoInfo*, pages 95–108.

Murray-Rust, P. (2013). Open Data in Science. *Serials Review*, 34(1):52–64.

Quoca, H. N. M., Quoca, H. N., Hauswirtha, M., and Le Phuoca, D. (2014). Global weather sensor dataset.

The Open Definition (2013). Open definition. Acessed in `http://opendefinition.org/od/2.0/pt-br/`.

Turner, B. L., Lambin, E. F., and Reenberg, A. (2007). The emergence of land change science for global environmental change and sustainability. *Proceedings of the National Academy of Sciences*, 104(52):20666–20671.

Wick, M. and Vatant, B. (2012). The geonames geographical database. *Available from World Wide Web: http://geonames. org*.