Linked Open Data in Smart Cities: An application in the domains of Mobility and Education

Mateus G. Belizario, Rita Cristina G. Berardi

[1]Programa de Pós-Graduação em Computação aplicada (PPGCA)
Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba – PR – Brazil

mateusbelizario@alunos.utfpr.edu.br, ritaberardi@utfpr.edu.br

Abstract. *The fast transition to a highly urbanized population means that governments face new challenges in managing data and information in urban spaces. One of them concerns the need for information models of similar semantics and the ability to share and connect information from different sources and in heterogeneous formats, allowing the improvement of operational decision making by managers and citizens. In this sense, the objective of this work is to carry out a proof of concept integrating open data from the domains of education and mobility, using ontologies. The data used are from the city of Curitiba relating the academic performance of students at different levels and the characteristics of urban mobility in the city. The linked data obtained is an indication of the relationship between the academic performance of students in the municipality and the characteristics of urban mobility in the city.*

1. Introduction

The fast transition to a highly urbanized population has led some governments to face new challenges in relation to key issues such as sustainable development, education, energy, the environment, security and public services, among others. The availability of Information and Communication Technologies (ICTs) in smart cities stimulates the development of new services and applications, and creates a more efficient environment for collaborative problem solving and innovation [Bolívar 2018].

Towards the development of smart cities, [Mellouli et al. 2014] declare that governments over the years have engaged in a movement to open data with open licenses and in formats that are easier to reuse. *Smart data* is an example of technology that can be used to enable the ability to strategically manage government data through its opening, distribution and structuring [Algemili 2016].

The main challenges in managing data in smart cities, as [Naphade et al. 2011], are the need for common information models and the ability to share information from various agents and institutions holding this data within a city. Also according to these authors, to guarantee visibility when managing services and infrastructures of smart cities, it is necessary to integrate data from different sources, each with its own sampling frequency, characteristics, formats and semantics.For example, information related to the flow and mobility of citizens is spread across many different institutions and domains, including transportation and urban planning. Creating and applying a unified information model makes it possible to obtain a more complete picture of urban activity and its

situation. Semantic Web technologies have the potential to provide the basis for new electronic services, assist in decision making and in the development of new solutions in urban ecosystems. The use of ontologies, *Linked Open Data* (LOD) and other semantic technologies open up new possibilities in smart cities, as it may combine information from various sources for purposes such as statistics, analysis, maps and publications, inform users when the information matches their interests and describe products and services more accurately [Bischof et al. 2014].

In a view of the heterogeneity of information and data sets arranged by smart cities, it is possible to use ontologies for data integration and connection. Ontology appears as a form of standardization and homogenization of data taking into account its semantics, in other words, the meaning of the data within its domain. Therefore, it is a data model used to represent the concepts, or classes, of a domain and its relationships [Guarino et al. 2009].

The problem addressed in this work is the lack of homogenization about the data format and its semantic values considering the context of open data in the city of Curitiba in the state of Paraná in Brazil. Specifically, two domains of the city were selected to work: Education and Mobility. These domains were selected because they demonstrate the difficulty of management, heterogeneity and integration of data available through the city's open databases.To show such difficulty consider that the citys educational institutions are located at certain addresses, and public transport lines run through certain streets and the bus terminals are located at specific addresses. However, this information is isolated and defined by different terms, therefore they are not connected.

Furthermore, according to [Heinlein and Shinn 2000], the influence of student mobility on schooling, cognitive development and academic performance has been studied. One of the relationships demonstrated that the greater the effort of commuting, the lower the student's achievements and income are during his academic life.

Establishing a relationship between mobility and educational achievement is a complex problem, however there are variables that can be considered. For example, studies show that children who move more than average to get to school are more likely to be poor, more likely to come from a home with a single parent, and are more likely to be in a home where the resident is unemployed or unable to graduate from high school [Long 1992].

2. Theoretical Foundation

For this research it was necessary to develop a framework of knowledge on the following issues: Smart cities, urban mobility and education, linked data, open data and ontological models.

In this work, the concept of smart city presented by [Giffinger and Pichler-Milanović 2007] will be adopted, which concerns different domains for the city and which covers both the technical and socio-cultural fields of the deployment of the use of technology in the active life of cities and their citizens. According to [Giffinger and Pichler-Milanović 2007], smart city is a city that performs well in the economy, in people, in governance, in mobility, in the environment and in life, built on the intelligent combination of resources and activities of self-confident, independent and conscious citizens.

Before relating urban mobility and education, it is important to understand the meaning of the two isolated terms. Mobility indices are difficult to define and quantify, but researchers include dimensions such as cause, distance, quantity, time and location. In addition, there is a distinction between school mobility and residential mobility, as a change of address does not require a change of school and vice versa. A student's school performance, on the other hand, is a more constant variable, since it can be measured by achievement tests, grades and grade-age progress, for example [Heinlein and Shinn 2000]. Urban mobility is composed of several modes, such as public transport, quality of roads and cycle paths, for example. The focus of this work is on the analysis of accessibility to public transport.

The connection of open data through the Internet as a Linked Open Data(LOD) offers the possibility of using data across domains or organizational boundaries for statistics, analyzes, maps and publications. By connecting this data, the interrelationships and correlations can be understood quickly. The added value is created when the stored data, not connected before, is combined and new conclusions can be reached [Geiger and Von Lucke 2012]. In this sense, the Semantic Web relies heavily on formal ontologies to structure data for a comprehensive and transportable understanding by machines. They serve as metadata schemes, providing a controlled vocabulary of concepts, each with an explicitly defined and machine-processable semantics [Maedche and Staab 2001]. They serve as metadata schemas, providing a controlled vocabulary of concepts, each with an explicitly defined and machine-processed semantics. By defining common and shared domain models, ontologies help people and machines to communicate concisely, supporting semantic exchange and not just the [Maedche and Staab 2001] syntax. In this context, computational ontologies are a means of formally modeling the structure of a system, that is, the relevant classes and relationships that emerge from its observation and that are useful for certain purposes [Guarino et al. 2009].

3. Ontology development and Data Connection

This section presents the process followed to develop the ontology and to link the data, it will be divided into the following subsections: Information needs, datasets, ontology developed and linking open data.

3.1. Information Needs

To understand the need for information by specialists in the two selected domains, questionnaires were developed and made available from October to December 2019. In total, twenty-eight people answered the form, nineteen students from higher education, eight specialists in the field of education and one specialist in the field of mobility.

The information needs raised were: Access to information on bus lines, number of bus stops and terminals close to educational institutions and on the academic performance of schools. These needs were essential to define the competence questions of this work that limit the scope of the ontology to be created [Noy et al. 2001b].

Based on responses from domain experts and students, four competency questions were developed that integrate data from both domains, education and urban mobility. They are: i) What are the lines next to a particular school?; ii) What are the access points

near certain school?; iii) What is the relationship between academic performance and access to public transport?; iv) What is the average mobility access point between schools with a specific Socioeconomic Level Indicator (INSE)[1]?

The first two competency issues were defined based on the needs that education experts demonstrated, especially concerns about student access and travel. The second two questions were inspired by the information needs demonstrated by the specialist in urban mobility. In other words, demonstrate the impact of mobility on the quality of education indexes and investigate whether access to public transport is different according to the region of the city, and the socioeconomic indicators of the educational institution.

3.2. Datasets

The first dataset used refers to Curitiba's open data on public transport stored on the EUBra-BigSea [2] in a PostgreSQL5 database. According to the creators of the project, [Alic et al. 2019], EUBra-BIGSEA is a collaboration aimed at developing cloud-based data analysis services adapted mainly for public transport data.

The EUBra-BigSea server also has data on education in the municipality of Curitiba, but it is intended to complement this data with the set of microdata made available in an open format by the National Institute of Educational Studies and Research (Anísio Teixeira) (INEP).

INEP microdata [3] provide information on the basic and higher education census of institutions from all over the national territory and on their performance in exams across the nation, such as Enem and Enade. For data referring to education, data from basic education, high school and higher education in the city of Curitiba were used.

In addition to these two datasets, the *Web Service* [4] from URBS were utilized, URBS its a mixed economy company that controls the public transport system in the city of Curitiba. This service provides *endpoints* for accessing specific information and data about Curitiba's public transport.

The data used for the proof of concept were the data for the year 2017, as they are the most updated and present in all datasets. In addition, with the exception of EUBra-BigSea data that was made available in a database, all files were in CSV (Comma-separated values) format.

3.3. Lightweight Ontology Developed

The development of this lightweight ontology was based on the processes established in *Ontology Development 101* de [Noy et al. 2001a]. As tools to support development, *software* Protégé in version 5.5.0 and a *plugin* called *OnTop* [5] for mapping data in relational format to triples. Table 1 shows the list of classes defined in the ontology. Fourteen classes were defined, four of which were reused from other ontologies.

[1]The INSE can vary from 1 to 6, with 6 being the highest level group, which indicates that students in general have at home a high quantity of elementary goods. Available at http://portal.inep.gov.br/web/guest/indicadores-educacionais

[2]Available at https://www.eubra-bigsea.eu

[3]Available at https://portal.inep.gov.br/web/guest/dados

[4]Available at http://transporteservico.urbs.curitiba.pr.gov.br/login.php

[5]Available at https://github.com/ontop/ontop

The *Ontology Development 101* establishes the following processes for developing an ontology: determining scope; consider reuse; enumerate terms; define classes; define properties; define restrictions and, create instances. These processes were grouped into two main activities: i) specification of the ontology, ii) acquisition of knowledge and iii) population of the ontology. The developed ontology aims to represent the main entities related to public transport, covering terms such as bus stops, bus lines and bus terminals, and education, covering basic, fundamental and higher education through assessment metrics. Figure 1 represents the relationships between the objects of different classes of the ontology. The Figure has a legend, where each *object property* is represented by a different color.

All files in *Comma-separated values* (CSV) format were handled and had values corrected, for example. In addition, special treatments were performed. Records with null values were removed from essential data, such as the INEP code for schools and universities, and the values of the Socioeconomic Level Indicator of Basic Education Schools (INSE), which in some CSVs were presented in text format and in others, were normalized. represented by whole numbers.

After processing the data, they were inserted in a PostgreSQL5 relational database, at the end the data were mapped in the ontology by the plugin *OnTop* in *Protegé*. In the final phase, the generated triples were inserted in a triple bank *GraphDB* [6], where the queries were held to validate the competence issues.

As a result of the ontology population, a file was obtained in textitturtle format (*.ttl*), with the triples generated by *plugin OnTop*. In general, 114,605 triples were created, and the file with the connected data is openly displayed on *Github*[7].

The results of this stage of development, the *.owl* file of the built ontology and the used ontologies, are openly displayed on *Github*[8].

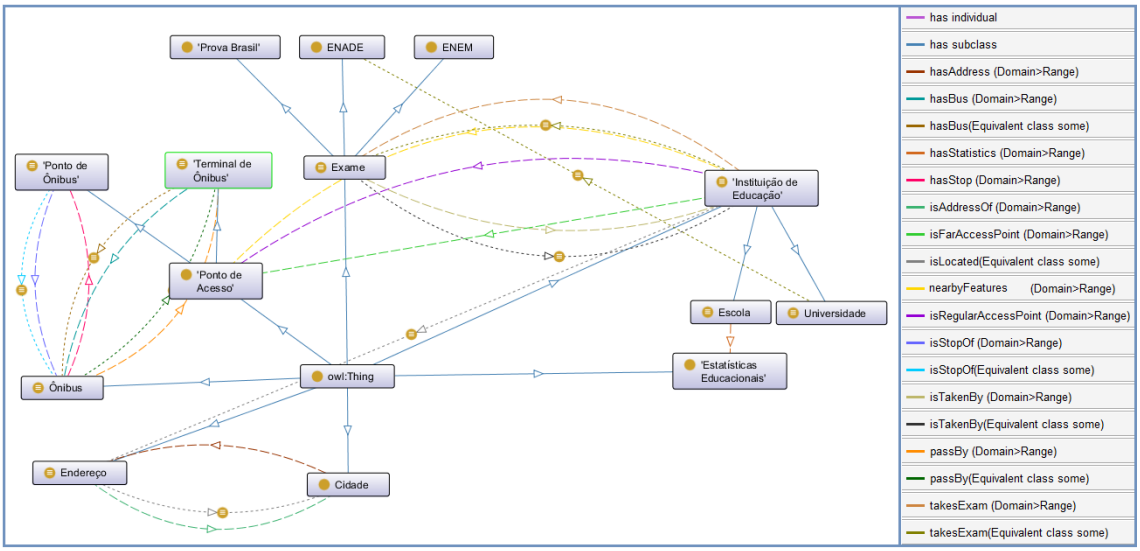

Figure 1. Class Relationship Map. Own authorship

[6]Available at: http://graphdb.ontotext.com/
[7]Available at: https://github.com/MateusBelizario/ontology_mobed_2020/tree/master/Dados_Conectados
[8]Available at: https://github.com/MateusBelizario/ontology_mobed_2020

Table 1. List of Classes

Terminological axiom	Description
Educational Statistics	The facts and figures on the quality of education for a given year.
City	Conglomerate of people who, located in a geographically delimited area, own many houses, industries, agricultural areas.
Address	Data needed to locate a property (street name, house number, apartment, floor, land, etc.)
Exam	Evaluation of the academic performance of an educational institution.
ENADE	Subclass of the class *Exam*, represents the National Exam of Student Performance.
ENEM	Subclass of the class *Exam*, represents the National Exam of High School Performance.
Prova Brasil	Subclass of the class *Exam*, represents the National Assessment of School Performance of elementary school.
School	Educational establishment, public or private, represents all educational institutions.
Secondary School	Subclass of the class *School*. Represents an educational institution focused on elementary and high school.
University	Subclass of the class *School*. Represents an educational institution covering higher education , postgraduate, master's, doctorate etc.
Access Point	Physical passage that allows use of public transport.
Bus Stop	Subclasse de *Access Point*. It represents the stop of a bus on the streets of the city on its way.
Bus Station	Subclass of *Access Point*. Structures where city buses stop for passengers to board and/or disembark.
Bus	Represents any vehicle that performs the function of transporting passengers in an urban area.

4. Results

In this section, the results will be exposed by showing just the results of the SPARQL queries due space limitation, however the complete group of queries and results can be accessed in a Github repository[9]. All SPARQL queries were made in the *GraphDB*.

In order to answer the competency question Q1: *What are the lines next to a particular school?* we choose a well known school DOM BOSCO in a known neighborhood called Batel that is a central neighborhood of Curitiba. A logical union is made between the bus lines that pass close to the school DOM BOSCO, through the bus stops and through the nearby bus terminals for the DOM BOSCO school from the neighborhood called Batel, with the identification code 2246. Figure 2 shows the result, the list with the name of the bus lines that pass within a radius of 500 meters from the school.

[9] Available at: https://github.com/MateusBelizario/ontology_mobed_2020

In order to answer the second competency question Q2: *What are the access points near a particular school?* we also choose the school DOM BOSCO. In it, access points near the school DOM BOSCO are selected, with identification code 2246. Figure 3 shows the result, it lists the school and the name of the nearby access points, which are the address of the stops from the terminal, if there were any.

	busName
1	CAMP.SIQ./BATEL
2	JD.SOCIAL/BATEL
3	RUA XV / BARIGUI
4	TRAMONTINA
5	V. SANDRA
6	AHÚ/LOS ANGELES

Figure 2. Result: Competency Question 1. Own authorship

As can be seen see in Figure 3, there are three bus stops located within 500m of this school, their address is indicated in the column *accessName*. The ontology also allows research on which buses go through these points, if it is of interest to managers and citizens, using the defined *isStopOf* property.

	schoolName	accessName
1	Dom Bosco - Batel	Rua Gonçalves Dias. 509 - Batel
2	Dom Bosco - Batel	Rua Bispo Dom José. 130 - Batel
3	Dom Bosco - Batel	Av. Sete de Setembro. 6001 - Batel

Figure 3. Result: Competency Question 2. Own authorship

Those results in Figure 2 and Figure 3 show some clear benefits of integration, as all databases were mapped in the ontology, it is possible to perform queries between data from the datasets, so it is possible to identify the bus lines (data from the EUBra-BigSea base) from the addresses of the schools in the municipality (data from the INEP database). In addition to using domain terms and not database identifiers as the data is usually stored.

In order to answer the third question of competence Q3: *What is the relationship between academic performance and access to transport public?* we choose the academic performance at the level of High School which is determined by ENEM data, although the data connection was made on several levels of education institutions.

Figure 4 shows the result, where schools are being ranked by the grade of the natural science discipline. The result shows only the five schools with the lowest scores in natural sciences, but originally the *query* brings all schools ordered. The results of the *query* do not clearly show the direct relationship between the domains, but it does provide the data necessary to carry out more complete investigations using statistical analysis that can gelp to clearify how mobility impacts the educational domain.

	name	accessPointNumber	natural	human	portuguese	essay
1	São Sebastião	"2"^^xsd:integer	"4,5306E2"^^xsd:double	"5,466E2"^^xsd:double	"4,8111E2"^^xsd:double	"5,6714E2"^^xsd:double
2	João Mazzarotto	"3"^^xsd:integer	"4,6214E2"^^xsd:double	"5,3317E2"^^xsd:double	"5,0363E2"^^xsd:double	"5,1931E2"^^xsd:double
3	Paulo Leminski	"3"^^xsd:integer	"4,6441E2"^^xsd:double	"5,4832E2"^^xsd:double	"5,0708E2"^^xsd:double	"5,3774E2"^^xsd:double
4	Aníbal Khury Neto	"7"^^xsd:integer	"4,6477E2"^^xsd:double	"5,5499E2"^^xsd:double	"4,9744E2"^^xsd:double	"5,0549E2"^^xsd:double
5	La Salle	"2"^^xsd:integer	"4,6509E2"^^xsd:double	"5,5545E2"^^xsd:double	"4,9985E2"^^xsd:double	"5,2272E2"^^xsd:double

...

Figure 4. Result: Competency Question 3. Own authorship

In order to answer the last competence question Q4: *What is the average mobility access point among schools with a specific INSE?* we choose to query the average number of access points for the schools with the INSE belonging to the 5 group.

Figure 5 shows a graph constructed with the results obtained from the query performed to the fourth competency question. In this graph, the averages for each discipline present in the ENEM test were plotted, grouped by the number of access points near the educational institutions. In general, the averages improved slightly even for institutions with five access points.

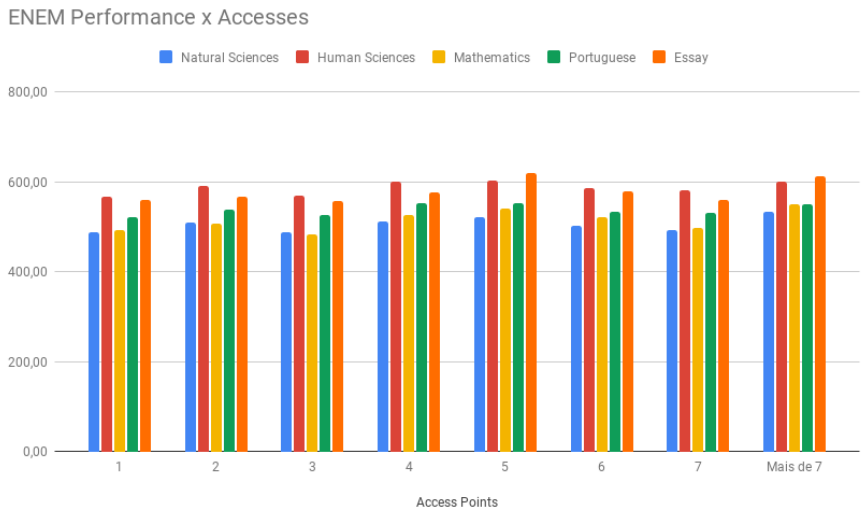

Figure 5. Average Performance in ENEM by Nearby Access Points. Own authorship

It was noticed that the highest scores of each ENEM subjects, with few exceptions, improved as the number of access points near the educational institutions increased until there were six access points to public transport. These data explain the relationship between the domains and may indicate that access to urban mobility can be a factor that contributes to improving academic performance.

To further explore the connection of the data, it was decided to carry out two more consultations: The amount of access to public transport close to educational institutions by neighborhood of Curitiba and the average number of access points to public transport by educational institutions by type of institution (private, state public or municipal public).

It was made a query to to carry out the consultation on the amount of access to public transport near educational institutions by neighborhood of Curitiba. As can be seen in Figure 6, the result for the *query* is shown, the five neighborhoods with the most access points near educational institutions are respectively: Cidade Industrial de Curitiba (CIC) largest neighborhood in Curitiba, Sítio Cercado, Centro, Uberaba and Portão.

	neighborhood	accessNumber
1	Cidade Industrial de Curitiba	"201"^^xsd:integer
2	Sítio Cercado	"103"^^xsd:integer
3	Centro	"94"^^xsd:integer
4	Uberaba	"85"^^xsd:integer
5	Portão	"76"^^xsd:integer

Figure 6. Number of Accesses to Public Transport by Neighborhood. Own authorship

A last consultation was made to query about the average number of access points to public transportation of educational institutions by type of institution (private, state public or municipal public). The results showed that the average number of accesses by type of institution is very similar. Private institutions have an average of 4.63 access points to public transport within a 500m radius, municipal public institutions have an average of 4.53 access points and state public institutions have an average of 4.32 access points to nearby public transport.

5. Conclusion

This work proposed the semantic integration of two domains of the urban space, specifically mobility and education, covering concepts and solutions *Linked Open Data*. Four specific objectives have been outlined to address this problem. By integrating open data on the domains, it was possible to obtain the public transport lines that are available within a radius of 500 meters from the school, discover the bus stops and terminals close to the school, perform a *query* to find out the average grades in academic performance assessments for the school and the average access points to public transport according to the educational institution's INSE. Through the answers of the competency questions we have showed the facilities offered by this kind of data when the objective is to integrate data from different sources in smart cities since to answer those competency questions without the semantic integration would be much more difficult.

As future work it is intended to extend the ontology to an ontological model that is capable of interpreting and connecting more domains of smart cities. It is also intended to explore ways to simplify the technical level necessary to consult information on the connected data, so that citizens can also develop solutions based on the data provided by this ontology. In addition carry out statistical analysiss on the new information on the relationship between the domains of education and mobility. This way, it is possible to propose new methodologies and technologies for citizen participation, which is a proposal aligned with chapter 6 of Grand Research Challenges in Information Systems in Brazil [Boscarioli et al. 2017].

References

Algemili, U. A. (2016). Outstanding challenges in recent open government data initiatives. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 6(2):91.

Alic, A. S., Almeida, J., Aloisio, G., Andrade, N., Antunes, N., Ardagna, D., Badia, R. M., Basso, T., Blanquer, I., Braz, T., et al. (2019). Bigsea: A big data analytics platform for public transportation information. *Future Generation Computer Systems*, 96:243–269.

Bischof, S., Karapantelakis, A., Nechifor, C.-S., Sheth, A. P., Mileo, A., and Barnaghi, P. (2014). Semantic modelling of smart city data.

Bolívar, M. P. R. (2018). *Smart Technologies for Smart Governments*. Springer.

Boscarioli, C., Araujo, R., and Suzana, R. (2017). Grand research challenges in information systems in brazil 2016-2026. *Brazilian Computer Society*.

Geiger, C. P. and Von Lucke, J. (2012). Open government and (linked)(open)(government)(data). *JeDEM-eJournal of eDemocracy and open Government*, 4(2):265–278.

Giffinger, R. and Pichler-Milanović, N. (2007). *Smart cities: Ranking of European medium-sized cities*. Centre of Regional Science, Vienna University of Technology.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer.

Heinlein, L. M. and Shinn, M. (2000). School mobility and student achievement in an urban setting. *Psychology in the Schools*, 37(4):349–357.

Long, L. (1992). International perspectives on the residential mobility of america's children. *Journal of Marriage and the Family*, pages 861–869.

Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2):72–79.

Mellouli, S., Luna-Reyes, L. F., and Zhang, J. (2014). Smart government, citizen participation and open data. *Information Polity*, 19(1, 2):1–4.

Naphade, M., Banavar, G., Harrison, C., Paraszczak, J., and Morris, R. (2011). Smarter cities and their innovation challenges. *Computer*, 44(6):32–39.

Noy, N. F., McGuinness, D. L., et al. (2001a). Ontology development 101: A guide to creating your first ontology.

Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., and Musen, M. A. (2001b). Creating semantic web contents with protege-2000. *IEEE intelligent systems*, 16(2):60–71.