

Alan Freihof Tygel

**Advancing on Use of Open Data Thorough Data
Literacy and Semantic Enhancement of
Metadata**

Brasil

Maio de 2016, v0

Alan Freihof Tygel

Advancing on Use of Open Data Thorough Data Literacy and Semantic Enhancement of Metadata

Tese de Doutorado submetida ao Corpo Docente do Programa de Pós-Graduação em Informática do Instituto de Matemática e Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Doutor em Informática.

Universidade Federal do Rio de Janeiro

Instituto de Matemática & Instituto Tércio Pacitti de Aplicações e Pesquisas
Computacionais

Programa de Pós-Graduação em Informática

Supervisor: Maria Luiza Machado Campo

Co-supervisor: Sören Auer

Brasil

Maio de 2016, v0

Alan Freihof Tygel

Advancing on Use of Open Data Thorough Data Literacy and Semantic Enhancement of Metadata/ Alan Freihof Tygel. – Brasil, Maio de 2016, v0-122 p. : il. (algumas color.) ; 30 cm.

Supervisor: Maria Luiza Machado Campo

Tese (Doutorado) – Universidade Federal do Rio de Janeiro
Instituto de Matemática & Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais
Programa de Pós-Graduação em Informática, Maio de 2016, v0.

1. Palavra-chave1. 2. Palavra-chave2. 2. Palavra-chave3. I. Orientador. II. Universidade xxx. III. Faculdade de xxx. IV. Título

Alan Freihof Tygel

Advancing on Use of Open Data Thorough Data Literacy and Semantic Enhancement of Metadata

Tese de Doutorado submetida ao Corpo Docente do Programa de Pós-Graduação em Informática do Instituto de Matemática e Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Doutor em Informática.

Trabalho aprovado. Brasil, 24 de novembro de 2012:

Maria Luiza Machado Campo
Orientador

Professor
Convidado 1

Professor
Convidado 2

Brasil
Maio de 2016, v0

Resumo

A publicação extensiva de dados em formatos abertos na Internet parece ser uma tendência irreversível. No caso de governos, pressões da sociedade por mais transparência vem fazendo com que cada vez mais, administrações públicas facilitem acessos aos dados de governo criando centrais únicas de acesso. Estes repósitorios centrais de dados, ou portais de dados abertos, têm como objetivo ser um ponto de encontro entre o governo e cidadãos e cidadãs que desejam acessar dados públicos. Espera-se com isso uma maior transparência das administrações públicas, que por sua vez alimenta a democracia com a população bem informada, e inibe desvios de recursos através da possibilidade de uma inspeção aberta ao público. Junto às grandes expectativas geradas pelas políticas de dados abertos, verifica-se também uma ampla gama de problemas que ainda retardam o crescimento das iniciativas. No decorrer da pesquisa relacionada a esta tese, dois problemas chamaram a atenção: (i) a falta de organização conceitual dos portais de dados abertos e (ii) as dificuldades do público em geral em lidar com os dados. Deste modo, esta tese busca trazer uma contribuição para o campo dos dados abertos propondo abordagens para ambos os problemas. Administradores dos portais de dados abertos utilizam diversos tipos de metadados para organizar seus conjuntos de dados, sendo o mais importante as tags. Entretanto, o processo de tageamento é sujeito a diversos problemas, como sinonímia, ambiguidade ou incoerência, entre outros. Como demonstrado por nossa análise empírica dos ODPs, estes problemas são atualmente prevalente na maior parte dos portais, e efetivamente dificultam o reuso dos dados abertos. Face a estes problemas, nesta tese foi desenvolvida e implemenatada uma abordagem para reconciliação de tags em portais de dados abertos, incluindo ações locais relacionadas individualmente aos portais, e ações globais para adição de uma camada de metadados semânticos acima dos portais individuais. Diversos estudo apontam ainda que, dentro das políticas de dados abertos adotadas pelos governos, há um foco maior na publicação de dados, e uma menor atenção à capacidade de uso destes dados pelo público. Assim, ainda que os dados sejam publicados, é necessário que haja um público capacitado para lidar eles. Do contrário, corre-se o risco de criar uma elite capaz de tirar proveito destas informações, e aprofundar ainda mais a exclusão digital, sobretudo em países com o Brasil. Neste sentido, apresentamos nesta tese uma abordagem para alfabetização em dados, inspirada na pedagogia da educação popular e na técnica de pesquisa-ação participativa. Deste modo, espera-se com essa tese contribuir com o avanço da democratização das informações, contextualizando de forma mais adequada a publicação de dados abertos, e permitindo um uso mais ampliado pela população.

Palavras-chave: Dados Abertos. Portal de Dados Abertos. Conciliação de Metadados. Enriquecimento Semântico. Alfabetização em Dados.

Abstract

The extensive publishing of data in open formats on the Web seems to be an irreversible tendency. Regarding governments, claims for more transparency coming from the civil society are forcing public administrations to ease the access to government data, mostly through central data repositories. These data repositories, or Open Data Portals (ODPs), have the objective of being a meeting point between government and citizens willing to access public data. Hence, it is expected a greater transparency of public administrations, which in turn feed democracy with a well informed population, and inhibits corruption through the possibility of open scrutiny by the public. Alongside the great expectations created by the open data policies, we also verify a wide range of problems which still hinder a more effective growing of the open data initiatives. During the research related to this thesis, two problems called the attention: (i) the lack of conceptual organization of the open data portals, and (ii) the difficulties of the general public for dealing with data. Thus, this thesis expects to bring a contribution for the field of open data by proposing approaches for both problems. ODP managers use several types of metadata to organize the datasets, one of the most important ones being the tags. However, the tagging process is subject to many problems, such as synonyms, ambiguity or incoherence, among others. As our empiric analysis of ODPs shows, these issues are currently prevalent in most ODPs and effectively hinders the reuse of Open Data. In order to address these problems, we develop and implement an approach for tag reconciliation in Open Data Portals, encompassing local actions related to individual portals, and global actions for adding a semantic metadata layer above individual portals. Several studies attest that, on the open data policies adopted by governments, there is a greater attention on data publishing, while enhancing the capacity of people for using these data remains in background. Thus, even if data are published, it is necessary to have an empowered society to deal with it. Otherwise, there is a risk of creating an elite able to profit from these information, deepening even more the digital divide, especially in countries like Brazil. In order to tackle this matter, we present in this thesis an approach for data literacy, inspired in the pedagogy of popular education and in the participatory action-research technique. It is expected that this thesis contribute with and advance in the democratisation of information, contextualizing in a more adequate form the publication of open data, and allowing its use by a broader part of the population.

Keywords: Open Data. Open Data Portal. Metadata Reconciliation. Semantic Lifting. Data Literacy.

List of Figures

Figure 1 – Data Spectrum as a definition of steps between open and close data. . .	24
Figure 2 – Model to analyse open budget initiatives.	27
Figure 3 – Different uses of data, with process, summary and examples.	32
Figure 4 – Critical Data literacy process.	47
Figure 5 – Trade-off between interpretation autonomy and software skills needed. .	54
Figure 6 – Classification tree for open data engagement actions.	62
Figure 7 – Tagging Ontology (KNERR, 2006)	67
Figure 8 – MUTO Ontology	68
Figure 9 – SRTag RDF schema (LIMPENS; GANDON; BUFFA, 2013)	72
Figure 10 – Re-use of tags inside a portal. The graphic shows the distribution of the percentage of tags used only once.	74
Figure 11 – Distribution of the average number of tags used per dataset in Open Data Portals.	75
Figure 12 – Proportion of similar tags in ODPs, where the difference lies only capitalization or special characters.	76
Figure 13 – Relevant elements of the Semantic Tags for Open Data Portals system.	84
Figure 14 – Overview of the StodAp approach.	85
Figure 15 – Overview of the local tag processing.	86
Figure 16 – Overview of the local tag processing.	90
Figure 17 – Example of the STODaP model showing relationships of the global tag < http://stodap.org/tags/Budget >.	92
Figure 18 – Local tag curation in a CKAN instance.	96
Figure 19 – Detail of dataset in an ODP.	97
Figure 20 – Semantic Tag Server for Open Data Portals.	97
Figure 21 – Correspondence between local and global tags. The yellow bar shows the number of exact occurrences of the tag in ODPs. The red bar shows the improvement when considered translations and synonyms, which can also occur in a same portal. This explains the numbers over 90. . .	98

List of Tables

Table 1 – Decontextualized phrases used in traditional literacy method, in Brazil.	42
Table 2 – Relation between Freire’s Literacy Method and data literacy.	45
Table 3 – Examples of data driven statements used to stimulate a critical view of data sources (based on Brazilian statistics agencies)	52
Table 4 – Open and closed analogies to help understand what open data is.	53
Table 5 – Examples of society driven databases, used by social movements with several purposes.	55
Table 6 – Summary of the presentations of the open data course for social movements.	57
Table 7 – Questionnaire answered by course attendants.	58
Table 8 – Summary of data used in the experiment.	73
Table 9 – Expressiveness of tags.	78
Table 10 – Examples of tags in each step of the procedure.	88
Table 11 – Examples of groups in some ODPs	89
Table 12 – STODaP evaluation - summary of subjects profile	101
Table 13 – Motivations, Impediments and Improvements indicated in answers to Question 4.	120
Table 14 – Impediments pointed in answers to Question 8.	121
Table 15 – Improvements indicated in answers to Question 9.	122

List of abbreviations and acronyms

CSO	Civil Society Organization
FOIA	Freedom of Information Act
ICT	Information and Communications Technology
LOD	Linked Open Data
NGO	Non-Governmental Organization
ODP	Open Data Portal
OGD	Open Government Data
OGP	Open Government Partnership
STODaP	Semantic Tags for Open Data Portals

Contents

1	INTRODUCTION	15
1.1	Motivation	15
1.2	Not Only Advantages – Open Data Impediments	16
1.3	Hypothesis	17
1.4	Objective	17
1.5	Methodology	17
1.6	Structure of the thesis	18
2	OPEN DATA – AN OVERVIEW	19
2.1	Motivation: Why Open Data?	19
2.2	Historical Notes	21
2.3	Definitions	22
2.4	The Open Data Landscape	25
2.5	Open Budget Data	26
2.6	Evaluating Open Data Impacts	29
2.7	Open Data Value	29
2.8	The Problems of Open Data	30
2.8.1	The Missing Focus on Use of Open Data	30
2.8.2	The Problem of Open Data Capacities and Data Divide	31
2.9	Linked Data towards Semantic Organization of Open Data	33
2.10	Conclusion	35
3	OPEN DATA RESEARCH THROUGH DATA LITERACY	37
3.1	An Overview on Data Literacy	38
3.1.1	Data Literacy and Popular Education	40
3.2	Contributions of Paulo Freire for a Critical Data Literacy	41
3.2.1	Paulo Freire, Literacy and Popular Education	41
3.2.1.1	Investigation Stage	42
3.2.1.2	Thematisation Stage	43
3.2.1.3	Problematisation Stage	43
3.2.1.4	Systematisation Stage	44
3.2.2	Parallels between Literacy Education and Data Literacy	44
3.2.3	A Freirean Inspired Critical Data Literacy	45
3.2.3.1	The Emancipatory Character of Data Literacy	46
3.2.3.2	Data Literacy Process	46
3.2.3.3	Data Literacy Stages	47

3.2.3.4	Definition	50
3.2.4	Conclusions	50
3.3	Teaching Open Data for Social Movements: action and research for open data engagement	51
3.3.1	First Stage – Introduction	52
3.3.2	Second Stage – Data Sources	53
3.3.3	Third Stage – Tools	56
3.3.4	Fourth Stage – Final Work	56
3.4	Open Data Clues from the Field	57
3.4.1	Questionnaire based analysis	58
3.4.2	Observation based analysis	59
3.4.3	Synthesis	61
3.5	Conclusion	63
4	SEMANTIC METADATA FOR OPEN DATA DESCRIPTION	65
4.1	A Literature Review on Semantic Metadata	65
4.1.1	Characterization of the Contribution	68
4.1.2	Metadata Assessment	69
4.1.3	Metadata Clean-up	69
4.1.4	Metadata Reconciliation	70
4.1.5	Structure Emergence	71
4.1.6	Automatic semantic tagging	72
4.1.7	Semantic Lifting in ODPs	72
4.2	An analysis of metadata in ODPs	72
4.2.1	Local Metrics	74
4.2.1.1	Tag Reuse	74
4.2.1.2	Tags per dataset	75
4.2.1.3	Tag similarity	75
4.2.2	Global Metrics	76
4.2.2.1	Coincident tags between portals	76
4.2.2.2	Tag expressiveness	77
4.3	Our contribution regarding the state of the art	78
5	SEMANTIC TAGS FOR OPEN DATA PORTALS	81
5.1	Motivation	81
5.2	Definition of an Open Data Portal	83
5.3	Overview of the STODaP Approach	84
5.4	Building the STODaP server	86
5.4.1	Local Processing - Clean Up and Reconcile	86
5.4.2	Global Processing - Interlinking Portals	88

5.5	Use and Maintenance of the STODaP server	90
5.5.1	Local Part - Cleaning up tags	90
5.5.2	Global Part - Semantifying Tags	91
5.5.3	STODaP Server - Interlinking Portals	92
5.6	Implementation	93
5.6.1	Building Global Tags	93
5.6.2	CKAN Tag Manager Plugin	93
5.6.3	CKAN Semantic Tags Plugin	93
5.6.4	Semantic Tag Server	94
5.7	Results	94
5.7.1	STODaP Server	94
5.7.2	Local Level	95
5.8	Conclusions	95
6	EVALUATION	99
6.1	Methodology evaluation background	99
6.2	Experimental Setup	100
6.2.1	Subjects	100
6.2.2	Tasks	101
6.3	Procedure	101
6.4	Results	102
6.4.1	Validation	102
6.4.2	Analysis	102
6.5	Conclusions	102
7	CONCLUSIONS	103
	BIBLIOGRAPHY	105
	APPENDIX	115
	APPENDIX A – LIST OF PUBLICATIONS	117
A.1	Peer-reviewed conferences	117
A.2	Peer-reviewed journals	117
A.3	Book chapters	117
	APPENDIX B – RESULTS OF OPEN DATA RESEARCH	119

1 Introduction

This thesis is essentially about open data. Given the growing importance of the topic, open data is discussed through several points of views. Although looking at it from a Computer Science perspective, it was not possible to skip political, social and economical aspects while discussing the topic. This work should thus be regarded as a multidisciplinary effort to contribute to a field that is heavily related to Computer Science, but far from being restricted to it.

This introductory chapter briefly exposes some motivations behind the topic of open data, highlighting selected problems observed in the literature. The hypothesis from where this thesis starts are posed, as well as the objectives that we aim to achieve with this work. We further explain the methodology used in order to develop this thesis, and finally describe the structure of the rest of this text.

1.1 Motivation

Current numbers about the open data scene leave no doubt about the central importance of this topic in contemporary society. The [Open Data Index](#) monitored in 2015 open datasets published by 122 countries all over the world, on topics related to budget, national statistics, procurements, maps and many others. Regarding the European landscape, the [Open Data Monitor](#) counts 173 open data catalogues in the continent, which sums an amount of 1472 GB of data.

The movement towards opening datasets is based in a series of access to information laws. According to the [right2info.org](#) platform, these are laws that “establish the right and procedures for the public to request and receive government-held information”. [Vleugels \(2012\)](#) presents a comprehensive list of 273 Freedom of Information Acts (FOIA), being 93 of national, 180 of sub-national and 3 of international scope. Even though the first occurrence of this kind of law dates from 1766, in Sweden, the vast majority of them were created after the year 2000.

It is no coincidence that 9 years later, United States and United Kingdom launched their Open Data Portals (ODPs), a one-stop-shop for publishing and consuming government data. Nowadays, several countries already implemented their ODPs, together with numerous states and municipalities. Universities and research centres are also joining strategies for putting data available on the Web. More than 1600 ODPs were surveyed by [OpenDataSoft](#). The potential of changing the very basis of democratical processes took the United Nations to start using the term *Data Revolution* to designate “the new world of data, a world in

which data are bigger, faster and more detailed than ever before” (Data Revolution Group, 2014).

If the numbers about open data scenario generate an enthusiastic hope that this movement will solve many problems of the society, an opposite direction claims that the promises are still from being realised, and also that there are some hidden threats that should be alerted.

1.2 Not Only Advantages – Open Data Impediments

Despite its recent popularity, Open Data and Open Data Portals still face significant impediments, as richly described by Zuiderwijk et al. (2012). The authors collected 118 socio-technical impediments for use of open data from interviews, workshops and literature. Some cited impediments were “absence of commonly agreed metadata”, “insufficiency of metadata”, “the lack of interoperability” and “difficulty in searching and browsing data”, showing that a great challenge for ODPs is the organization of data.

The open data organization challenge can be subdivided into two aspects: 1) structuring and organizing the datasets themselves and 2) providing well-structured and organized metadata for the datasets. The first aspect was, for example, tackled by approaches for semantic lifting of data by Ermilov, Auer e Stadler (2013a) and Ding et al. (2011b), who tried to build general strategies for putting large open government datasets in the Link Data cloud. For the standardized structuring metadata, the Data Catalog Vocabulary (DCAT)¹ (CYGANIAK; MAALI; PERISTERAS, 2010) was developed. However, the cross-portal metadata alignment and reconciliation can not be addressed by DCAT.

Besides the data organization challenge, another very much cited impediment to the use of open data is the lack of capacity of individuals and groups for dealing with open data. There has been a recently growing consensus on defining this individual and collective capacity for dealing with data under the concept of *Data Literacy*.

As observed by Bhargava e Ignazio (2015), one of the first mentions of the term *Data Literacy* called the attention for its importance on the context of evaluation of information, together with Information Literacy and Statistical Literacy. In 2004, Schield reinforced the important of teaching these three literacies for “students who need to critically evaluate information in arguments” (SCHIELD, 2004).

Although not mentioning directly the term Data Literacy, the above cited collection of impediments of open data (ZUIDERWIJK et al., 2012) dedicates a section for problems related to *understand ability*. Among them we find, for example, “Lack of skills and

¹ Available at <<http://www.w3.org/TR/vocab-dcat/>>

capabilities to use the data and” and “Lack of knowledge about how to interpret the data”, which relates directly to the topic of Data Literacy.

1.3 Hypothesis

In the light of the theoretical benefits of open data, and the impediments that hinder the achievement of these benefits, we formulate two hypothesis in order to guide the development of this thesis:

H1: Enhancing the organization of open data repositories leads to better use of open data;

H2: Increasing the level of data literacy on the society leads to better use of open data (which in turn motivates better publishing).

1.4 Objective

To develop approaches both from user perspective and from publisher perspective in order to advance towards the benefits of open data. Specifically:

- From the publisher perspective, and recognizing that the lack of organization of ODPs is a problem, to develop an approach for cleaning, reconciliation, and semantic lifting of metadata;
- From the user perspective, and recognizing the lack of abilities for dealing with data, to develop a data literacy approach

1.5 Methodology

The methodology used to develop this thesis is composed by several steps.

- Literature revision, described in chapter 2
- Participatory research, described in chapter 3
- Objective alignment
- New and deeper literature revision, described in chapter 4
- System development
- Validation: In order to validate our approach, experiments and objective metrics were developed. Specifically:
 - Hypothesis 1 was validated by measuring the ODP metadata related parameters;
 - Regarding Hypothesis 2, the developed data literacy method was applied and results were analysed.

1.6 Structure of the thesis

The remaining of this thesis is organized as follows:

Chapter 2 presents an introduction to the main subject of this thesis, i.e., open data. After an overview on the topic, we present some current research challenges in this field, with a special focus on the lack of people's capacity for dealing with data, and the lack of organization and linking possibilities on Open Government Data Portals. For some parts of this chapter, the work published in [Tygel et al. \(2016a\)](#) will be used.

Chapter 3 has a twofold objective: on the one hand, we present the results of a participatory research with open data users about their main motivations and impediments, and the wanted improvements on open data platforms. On the other hand, we systematise the research method into an open data course inspired in the principles of Popular Education. The course methodology is presented, as well as some contributions on critical data literacy. As a result of the research, the problem of linking and organizing data in ODPs appears as an outstanding impediment for using open data. In this chapter, we will use the results published in [Tygel, Campos e Alvear \(2015\)](#) and [Tygel e Kirsch \(2015\)](#).

Chapter 4 goes deeper in analysing how previous research dealt with enhancement of metadata in ODPs. A special focus is given on methods for extracting semantics of metadata, specially when dealing with tags.

Chapter 5 presents the STODaP – Semantic Tags for Open Data Portals – approach. The main purpose of this approach is to tackle the issue of OGD organization and linking, by cleaning up tags in ODPs, and creating a central repository for semantically annotated metadata. The approach is composed by several strategies, both in the context of individual ODPs and between them. As shown by our use case, ... In this section, we will benefit from the work published in [Tygel et al. \(2016b\)](#).

Chapter 7 will present the concluding remarks of this thesis, signalling ways for researchers willing to continue this work.

2 Open Data – An Overview

Open data is currently a very popular concept. As discussed in the previous chapter, it is part of an important political debate related to transparency, citizen participation, and considered a crucial way for improving democracies around the world. In this chapter, we drive an overview about open data, with the objective of historically contextualizing open data, highlighting the problems and challenges currently observed. Being a very dynamic field, it is impossible to picture the open data field only looking at academic works. Thus, the material used to write this chapter also includes websites, practitioners reports and official documents, seeking to reflect more clearly the current open data landscape. Rather than exhausting the topic, the aim of this chapter is to justify the importance of open data, present the open research issues and point the solutions to be presented in the following chapters.

The chapter starts with a section dedicated to the alleged motivations for opening data, including other contexts than government. In the sequel, some historical notes about the open concept and the open data term are presented, followed by a collection of open data definitions. Section 2.4 reviews the efforts to map the open data landscape using different assessment methods. In order to ground the discussion in a more concrete basis, in the following section we selected one special type – open budget data – to describe in a more detailed fashion. The chapter continues with two sections that seek to evaluate the results of open data efforts: impacts evaluation (Section 2.6) and value creation (Section 2.7). Of crucial importance is Section 2.8, where the problems of open data are analysed. This section is following by the presentation of the Linked Data approach, which is regarded as a way to overcome some of the mentioned impediments. We finally conclude this chapter pointing selected references to a broader understanding of open data.

2.1 Motivation: Why Open Data?

There are several motivations on why one should publish open data. When data is related to government, and thus called Open Government Data (OGD), reasons are even stronger, because it deals essentially with data related to public administration. According to the [Working Group on Open Government Data](#) at the Open Knowledge Foundation, there are three main motivations for governments to publish open data:

- Transparency;
- Releasing social and commercial value; and
- Participatory Governance.

The same organization curates a collaborative web book which presents a more extensive list of activities possibly benefiting from open government data ([Open Knowledge Foundation, 2015](#)):

- Transparency and democratic control;
- Participation;
- Self-empowerment;
- Improved or new private products and services;
- Innovation;
- Improved efficiency of government services;
- Improved effectiveness of government services;
- Impact measurement of policies; and
- New knowledge from combined data sources and patterns in large data volumes.

A comparison between open government data implementation strategies in 5 countries driven by [Huijboom e Broek \(2011\)](#) concluded that there are three primary motivation for governments to publish open data:

- Increasing democratic control and participation;
- Foster service and product innovation; and
- Strengthen law enforcement.

Although very important, government data is not only kind of data possible to be opened. Another important field where open data is discussed is science. According to [Murray-Rust \(2008\)](#), copyright over scientific data “is a major impediment to the progress of scholarship in the digital age.” His work severely criticizes publishers who impose barriers to free use of academic papers and associated supporting information, such as datasets, experiments data, simulation source code or software output. The author strongly defends an Open Access policy for publishing scientific work, and also lists a number of reasons why scientific data should be open:

- “Data belong to the human race.” Typical examples are genomes, data on organisms, medical science, environmental data;
- Public money was used to fund the work and so it should be universally available;
- It was created by or at a government institution (this is common in US National Laboratories and government agencies)
- Facts cannot legally be copyrighted;
- Sponsors of research do not get full value unless the resulting data are freely available;
- Restrictions on data re-use create an anticommons;
- Data are required for the smooth process of running communal human activities (map data, public institutions); and
- In scientific research, the rate of discovery is accelerated by better access to data.

2.2 Historical Notes

Although the idea was present in the scientific world for a long time, the term open data appeared for the first time in 1995, regarding the opening of geophysical and environmental data in an American scientific agency (CHIGNARD, 2013). Tauberer (2014) also affirms the roots of open data praxis come from the scientific community, who first realized the importance of opening and sharing data. He affirms that open government data, in turn, “has its own history rooted in Web 2.0, political campaigns, and innovations inside of municipal governments.”

The open source software movement fights since the 1980’s for the source code of software to be open and free¹. However, with the popularization of the Web, the increased speed in transmission rates, and the widely spread concept of Web Application, it was recognized that opening the source code was not enough for the unrestricted flow of knowledge through the Web. It was necessary that, beyond the code, public data could also be open, and also considered a common good, a thus not subject to private appropriation.

According to Chignard (2013), in 2007, a meeting between thinkers and activists in Sebastopol, USA, defined some concepts about open data, and some strategies in order to effectively apply it. The basic idea is that public data are of common property, as well as in the scientific world.

The first days of year 2009 watched the release of a Memorandum on Transparency and Open Government (OBAMA, 2009) by the newly elected administration of Barack Obama, in the USA. The memorandum is a political commitment on transparency, public participation, and collaboration, stating that “Openness will strengthen our democracy and promote efficiency and effectiveness in Government”. On the same year USA and UK released their open data portals in order to centralize the distribution of open government data. This action was followed by several countries and local administrations, as we will see in Section 2.4.

The first academic papers about open data started to be published only in 2010, according to a survey driven by Attard et al. (2015). One year, in 2011 the Open Government Partnership (OGP) was launched by eight countries, aiming to be a platform for national governments to willing to be more open, accountable, and responsive². In 2015, 69 countries were taking part on it and implementing their 1st, 2nd or 3rd action plans.

Another important historical milestone was the signature of the Open Data Charter³ by the G8 leaders, in 2013. The charter is based on six principles to be followed the

¹ Richard Stallmann always remembers that “free” has the sense of “free speech”, and not the sense of “free beer”. However, we must remember that a free beer in the sense of free speech (where the recipe is freely shared) also exists, available at <<http://freebeer.org/>>.

² Available at <<http://www.opengovpartnership.org/>>

³ Available here: <<http://opendatacharter.net/>>

governments that adopt it:

- Open by Default;
- Timely and Comprehensive;
- Accessible and Usable;
- Comparable and Interoperable;
- For Improved Governance and Citizen Engagement; and
- For Inclusive Development and Innovation.

In 2014, the charter was launched to the G20 group, and in 2015, it was also discussed at the Climate Conference, in Paris. According to the Open Data Charter portal, only a few countries already adopted it: Mexico, Uruguay, Chile, France, Italy, UK, Philippines, Guatemala and South Korea.

2.3 Definitions

One of the most used and accepted definitions of OGD are the Eight Principles of Open Government Data⁴, published as a result of the 2007 Sebastopol experts meeting. The eight principles are:

1. Complete: All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. Primary: Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. Timely: Data is made available as quickly as necessary to preserve the value of the data.
4. Accessible: Data is available to the widest range of users for the widest range of purposes.
5. Machine processable: Data is reasonably structured to allow automated processing.
6. Non-discriminatory: Data is available to anyone, with no requirement of registration.
7. Non-proprietary: Data is available in a format over which no entity has exclusive control.
8. License-free: Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

From this definition, it should be noted that several dimensions of data publishing are tackled. The first three principles are about the *nature of data*, i.e., aspects related to the content represented by data. The next three ones are about *access to data*, dealing with aspects that impact the technical use ability of data. Finally, the last two principles deal with *legal framework over data*. However, there is a possible ambiguity on the last principle. The term *License-free* can be understood both as *free of license*, i.e., there

⁴ Available at [<https://opengovdata.org/>](https://opengovdata.org/)

is no legal framework regulating what can and what cannot be done with data, or as possessing a *free license*, i.e., a defined legal framework which guarantees that data is open. Nowadays, there is a certain consensus that the latter interpretation is the most productive, since it gives legal parameters for people wanting to re-use data, including for commercial purposes. Thus, some countries defined their own Open Data Licenses, e.g. Germany⁵ and UK⁶. The Open Data Commons platform⁷ offers legal support for open data and defines three types of license: Public Domain Dedication and License (PDDL), Attribution License (ODC-By), and Open Database License (ODC-ODbL).

To these 8 principles, another 6 ones were added by Tauberer (2014):

9. Permanent: Data should be made available at a stable Internet location indefinitely.
10. Safe file formats: “Government bodies publishing data online should always seek to publish using data formats that do not include executable content.”
11. Provenance and trust: “Published content should be digitally signed or include attestation of publication/creation date, authenticity, and integrity.”
12. Public input: The public is in the best position to determine what information technologies will be best suited for the applications the public intends to create for itself.
13. Public review
14. Interagency coordination

Another widely accepted definition comes from the general Open Definition, which is currently on version 2.1⁸. In contrast to the previous definition, this one is not aware from aspects related to the nature of data, probably because it was originally formulated for open source software, but is currently being applied for data and art works, among others. Access to data (e.g., Machine Readability and Open Format) and legal framework (Open License or Status) are covered by this definition. One advance of the Open Definition is the characterization of conditions that limits the open criteria, such as Attribution (require distributions of the work to include attribution of contributors, rights holders, sponsors, and creators), Integrity (modified versions of a licensed work should carry a different name or version number from the original work) and Share-alike (distributions of the work should remain under the same license or a similar license).

In order to help publishers in creating a roadmap towards open data, Berners-Lee (2010) proposed a five stars schema, where each star represents one step further in turning data more accessible. The scheme starts from a simple PDF file and finishes with the implementation of the Linked Open Data paradigm (see more Section 2.9). The key elements for each of the five stars are:

⁵ Available at <<https://www.govdata.de/dl-de/by-1-0>>

⁶ Available at <<http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>>

⁷ Available at <<http://opendatacommons.org>>.

⁸ Available at <<http://opendefinition.org/od/2.1/en/>>

1. Open License
2. Machine Readable
3. Open Format
4. Dereferenceable URIs
5. Linked Data

Though not so much cited, the Three Laws of Open Government Data developed by Eaves (2009) has the advantage of being written in a colloquial way, supposedly more accessible for non-experts:

1. If it can't be spidered or indexed, it doesn't exist
2. If it isn't available in open and machine readable format, it can't engage
3. If a legal framework doesn't allow it to be repurposed, it doesn't empower

Finally, the Open Data Institute defines a Data Spectrum, which ranges from closed, through shared until open data. Figure 1 pictures this definition.

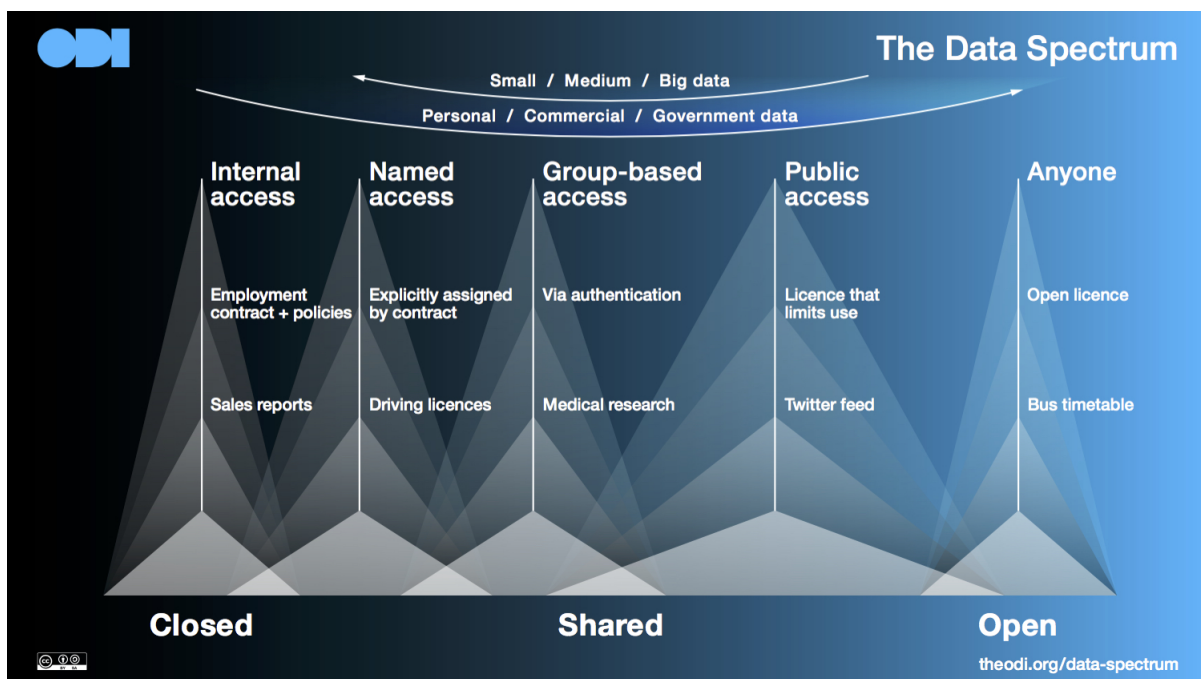


Figure 1 – Data Spectrum as a definition of steps between open and close data. Source: <https://theodi.org/data-spectrum>

With the same idea of defining intermediate levels of data openness, Bargh, Choenni e Meijer (2016) proposed the Semi Open-Data Paradigm. The objective is the analyse the dissemination of data through several dimensions, as publicity, completeness, timeliness, metadata, and other. For each dimension, several level should be defined. On the publicity dimension, the proposed levels are: ‘share with no one’, ‘share data within a specific group’, ‘share data within a department of an organization’, ‘share data within an organization’.

/ministry’, and ‘share data among a federation of organizations’ and finally ‘share with the public’.

2.4 The Open Data Landscape

While the number of open data initiatives around the world increases dramatically every year, several research projects driven from academy and/or civil society organizations seek to map the open data landscape. In the following, some of these projects are summarized, and their main results are presented:

Open Data Index: The Open Data Index is one of the most important platform for analysing the open data landscape in the world. In 2013, the first year, Open Data Index analysed 60 countries. In 2014 this number grew to 97, and in 2015 the evaluation covered 122 countries. The methodology consists basically in analysing datasets from 13 categories: National Statistics, Government Budget, Legislation, Procurement tenders, Election Results, National Map, Weather forecast, Pollutant Emissions, Company Register, Location datasets, Water Quality, Land Ownership and Government Spending. For each category, 9 features are evaluated with yes or no answers: “Openly licensed?”, “Is the data machine readable?”, “Is the data available for free”, “Available in bulk?”, “Is the data provided on a timely and up to date basis?”, “Is the data available online?”, “Is data in digital form?”, “Publicly available?” and “Does the data exist?”. From this analysis, a ranking is constituted according to each countries *score*. This score is a weighted sum that reflects the performance of each category for each feature. Considering all the countries, only 9% of the datasets are open. However, a strong inequality between the countries can be seen: while 25 of them have 50% or more datasets open, 44 have less than 25%.

Open Data Barometer: The Open Data Barometer also focus on a comparison of the open data context between countries. However, a more complex methodology is used to analyse each country, including expert interviews and secondary data, apart from accessing the datasets in a similar way as the Open Data Index. The 2nd Edition of this research, released in January 2015, analysed 86 countries concluded that “there is still a long way to go to put the power of data in the hands of citizens”(DAVIES; SHARIF; ALONSO, 2015) .

Open Data Monitor: The Open Data Monitor is focused on looking at datasets from European countries. One interesting aspect of this project is the measurement of “availability”, which considers the existence of “a description, at least one resource with a functional link and an available email of the author” for datasets in a catalogue. Surprisingly, the first three countries with more datasets (Germany, UK and Spain) have only a bit more than half of their datasets available (51%, 63%, 57%, respectively).

Open Data Inception: This project presents the largest geotagged listing of open

data portals, with more than 1600 ODPs showed in a map. For each portal, URL and associated geographical region is given.

Right2Info: This platform monitors FOIAs, which is not specifically open data, but is very related. 93 countries have some kind of FOIA, and the platform presents a comprehensive list of 273 FOIAs covering various scopes (VLEUGELS, 2012).

2.5 Open Budget Data

From all types of OGD, one is of particular importance: government budgetary data, as timely access to these data is critical to accomplish government accountability.

All governments and public administrations maintain budgetary data, unlike, for example, bus position data, which depends on sensors, or data about the occurrence of a specific disease, which depends on a health information system. From the citizen side, information on budget is a key element to ensure that public funds are being properly used. In locations where a participatory budget (MKUDE; PÉREZ-ESPÉS; WIMMER, 2014) was implemented, that is, part of the budget allocation is decided by the community, access to this kind of data is indispensable. A global initiative to improve openness of governments – the Open Government Partnership (OGP) – has the fiscal transparency as minimum eligibility criteria⁹, characterizing budget data as a foundation of open government.

Even with so many possible positive impacts, existing public financial transparency portals suffer from a number of shortcomings. First of all, they suffer from the large number of diverse data structures that make the comparison and aggregate analysis of transnational financial flows practically impossible. The tools to present, search, download and visualise this financial data are also nearly as diverse as the number of existing portals. This heterogeneity may even prevent an analysis of the quality of the data for the same funds administered by different funding authorities (VAFPOULOS et al., 2013). Past efforts have sought to overcome this situation by creating comprehensive and connected transparency portals, such as Farmsubsidy.org, and more recently, Publicspending.net. Within the existing open budget initiatives, low user engagement has been reported (WORTHY, 2013). Moreover, most of the budget publishing efforts results in simple data catalogues, fragmented and dispersed, because they do not share standards and methodologies (VAFPOULOS et al., 2013). The absence of standards can lead to data misuse (ZUIDERWIJK; JANSSEN, 2014), or even to results opposed to the initial aims (GURSTEIN, 2011).

In Tygel et al. (2016a), we proposed a *structured analysis framework* in order to explicitate problems generated by the lack of standards and help policy makers to understand the importance of various aspects of budget data publishing. We also envision

⁹ Other criteria can be found at <<http://bit.ly/1929F1l>>.

the framework as a tool to design more adequate budget publishing systems. Together with other ongoing initiatives (OPENSENDING, 2014; VLASOV; PARKHIMOVICH, 2014), we believe that the development of a solid standard can help governments to make their budget data more usable, and thus enable citizen participation in the democratic process. The framework can be seen in Figure 2.

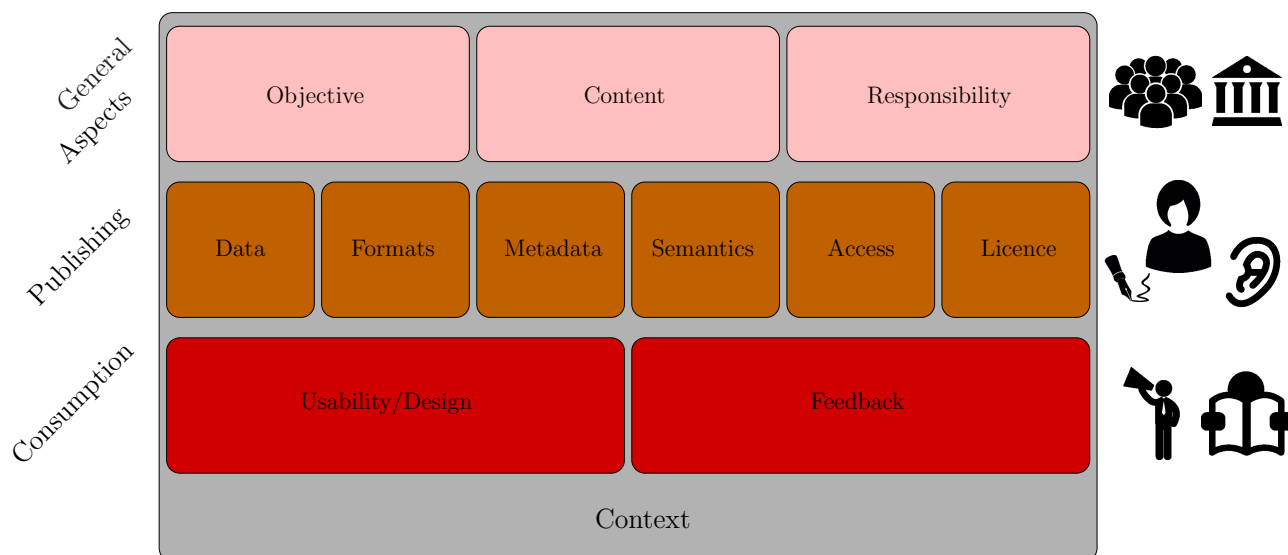


Figure 2 – Model to analyse open budget initiatives. The four parts – General Aspects, Publishing, Consumption and Context – are interconnected, and composed by several dimensions. Icons: Flaticon (CC).

Results from the application of this framework to 23 open budget initiatives can be seen at <http://bit.ly/1FNThhH>. The goal of the evaluation is not to be extensive or to achieve statistical significance, but rather to test the model, to discover its potentials and limitations, and to gain some intuition on the domain.

The 23 initiatives were chosen considering a balance between primary (11) and secondary (12) sources. The sample also contains at least five initiatives strongly related to each use perspective, and considers initiatives from 6 countries plus the European Union, presented in five different idioms. Some of the analysed initiatives are listed on the *Map of Spending Projects*¹⁰.

All primary sources are maintained by the government, and most of the secondary ones are society driven. Among them, two initiatives were identified as maintained in partnership between government and society organizations. Initiatives generally display their objectives (22), but only 11 explicitly mention their intended audience. Also, almost all initiatives offer data for download (18), which favours transparency perspective, and more than half of them (13) make visualization available, favouring participation perspective.

¹⁰ Available at <http://community.opensending.org/map-of-spending-projects/>.

Even considering the low number of initiatives evaluated, two outcomes drew the attention, regarding feedback and semantics. Commenting on data is allowed only in three initiatives, and the same number (but not the same ones) offers a data request form. No reporting issues mechanisms were found, revealing a strong absence of feedback possibilities.

The lack of semantics support (only three offered it), or linkable data (again, only three had it) also may point that policy marking perspective is still far from reality. Ten initiatives use categories for the datasets, which at least facilitate some form of comparisons. Regarding the use perspectives, we can state:

Transparency: The main requirements for this use perspective – data on transaction level, machine readable formats and aggregation levels – were accomplished by most of the open budget initiatives. However, much work is still to be done concerning the feedback handling. We can say that, for most of the analysed cases, stakeholders interested in auditing government and in translating data into more accessible formats are partially satisfied.

Participation: The requirements set for this use perspective enforced human readable formats that allow citizens without deep budget knowledge to understand data and to participate in discussions. Slightly more than half of the initiatives present graphics, which can help quick insights over data. Only three initiatives offer maps to visualize budget data, what is coherent to the low number of initiatives that include the location dimension (eight). Another aspect emphasized in this use perspective was the usability and design. Considering the already mentioned limitations on assessing this issue, we noticed that ten initiatives use standard open source software tools. Although this is not the most relevant factor regarding usability, the use of standard tools favours users dealing with several open budget initiatives. Moreover, as open source tools, the more initiatives using these tools, the better they can be developed.

Policy Making: The main requirements in this perspective were the use of common classifications, vocabularies and ontologies, and the possibility of linking data with other databases. As already mentioned, semantics support was mostly absent. Comparison tools, also important in this case, were found only in three of the initiatives. Thus, this use perspective is still far from being realised in most of the analysed initiatives. All these indicate that working on standard terminologies and common conceptualizations as suggested by OpenSpending ([OPENSPENDING, 2014](#)) is highly desirable.

The application of the model to 23 open budget initiatives made possible to derive several conclusions related to the specific use cases. However, it would be necessary a larger number of analysis and more iterations of the inductive-deductive approach in order be sure about the completeness of the model.

2.6 Evaluating Open Data Impacts

Although mapping initiatives is a very important way of assessing the quality of published data between countries, very little is known about the final effects of these policies. Almost a decade after the implementation in large scale of open data policies, researchers and practitioners start to pose the question: how to assess open data impacts on the society?

In order to tackle this question, a theoretical background to analyse the impact of OGD was developed by [Granickas \(2013\)](#). Impacts are divided into economic, political and social, and for each of them, possible implementation issues and impact metrics are deeply discussed. Recently, a working group was created to develop methods for assessing open data. In their first report ([CAPLAN et al., 2014](#)), a draft of a framework is proposed.

Finally, a recent report run a thorough review over evidences of impacts of fiscal openness ([RENZIO; WEHNER, 2015](#)). While recognizing that there is a literature gap on testing causal effects, the most rigorous studies found a relation between open budget initiatives and the desired outcomes.

An impact evaluation and comparison between almost 30 Brazilian government transparency portals, on several administration levels, is presented by [Beghin e Zigoni \(2014\)](#). The analysis was based on the 8 Open Government Principles evaluated for each portal by experts. Despite being a well defined and wide accepted model, these principles are quite general, and do not refer to specific characteristics of budget data. Moreover, they cover basically the publisher side.

2.7 Open Data Value

Another way of assessing open data impacts is through the analysis of *open data value*. Releasing social and commercial value is cited under the main motivations for governments to publish open data (see Section 2.1). Thus, it is necessary to understand the chain of value addition over data, and specifically what activities may add value for data. [Jetzek, Avital e Bjørn-Andersen \(2013\)](#) developed a conceptual model of OGD value generation, where enabling factors lead to value generation mechanisms which should finally release social and commercial value.

[Attard, Orlandi e Auer \(2016\)](#) proposed a Value Creation Assessment Framework, which profits from previous works, and extends some aspects. The framework walks through the complete Government Data Life Cycle Processes, namely: data creation, harmonisation, publishing, interlinking, exploitation and curation, and defines implementation and impact aspects related to each stage.

2.8 The Problems of Open Data

(ROSEIRA, 2016)

The vast majority of research about open data assumes that publishing public data in open formats will bring mostly good impacts. In this sense, two works from the same research group made an in depth research on the negative sides of open data.

In the first one, [Zuiderwijk et al. \(2012\)](#) analysed the socio-technical impediments that hinders the use of open data via literature review, interviews and workshops. As a result, 118 impediments were summarized in 10 categories: availability and access, find ability, usability, understand ability, quality, linking and combining data, comparability and compatibility, metadata, interaction with data provider and opening and uploading.

The second paper focuses on the possible negative effects that governments may face on opening data. [Zuiderwijk e Janssen \(2014\)](#) conducted several interviews with public servants and data archivists to find out which negatives effect they were concerned with. As a result, 16 negative effects were listed, for example: “risk of violating legislation by opening data”, “privacy can be violated unintentionally”, “misinterpretation and misuse”, “not citizens but others profit from open data” and “wasting resources to publish invaluable data”. It is interesting to note the question of data value also appearing here. In fact, methods for determining the value of data for users could help publishers selecting in which data should they put efforts.

Regarding the risk of privacy violation, two recent episodes attest that these concerns should really be taken into account. As reported by [Hern \(2014\)](#), the opening of taxi trips data by the city of New York allowed the identity of drivers to be discovered, and in some cases, even the passengers could be identified. In a similar situation, the release of film ranking data allowed not only the identity of users to be unveiled, but also their political orientation, religious views or sexual orientation. In both cases, the attempt to anonymise data failed.

[Parycek, Schöllhammer e Schossböck \(2016\)](#) interviewed

2.8.1 The Missing Focus on Use of Open Data

Another aspect revealed by [Zuiderwijk e Janssen \(2014\)](#) is the lack of informations interest about the actual use of open data: “The interviewees stated that they do not have much more information about how the open data have been used than the number of downloads”.

One of the few works dealing with this subject was written by [Davies \(2012\)](#). According to the author, “The gap between the promise and reality of OGD re-use cannot be addressed by technological solutions alone”. Thus, he raises the necessity of considering

human factors that affect the use or no use of data. In this paper, a Charter of Open Data Engagement is proposed, aiming to derive a parallel of the Five Stars of Open Data (BERNERS-LEE, 2010), but from the users point of view.

The five stars of open data engagement are (DAVIES, 2012):

- Be demand driven
- Put data in context
- Support conversation around data
- Build capacities, skills and networks
- Collaborate on data as a common resource

In the same work, Davies criticizes the so called “application fallacy”. According to him, the narratives about OGD assume that someone will develop an application to consume and visualize data. However, in his master thesis, Davies (2010) ran a survey with 55 instances using OGD from data.gov.uk, which revealed that in most of the cases facts are directly identified within datasets. Data is then used to base discursive reports, or to generate derivative datasets.






Davies (2010) describes five ways of using open data. Figure 3 shows the categories, and the number of cases gathered on the survey.

As a contribution for future research, author cites some challenges in the social and technical fields. The priority, according to him, is to understand the process that occurs between data publishing and its use in a determined application. Through this understanding it will be possible to overcome the barriers for use of data. Moreover, it is necessary to explore the existent political structures, so that the informations brought by data can effectively generate social changes. Finally, the broader challenge is to better understand the user point of view. The greatest technical challenge associated is to create tools that not only show data, but that support discussion and interaction around them.

2.8.2 The Problem of Open Data Capacities and Data Divide

Still according to Zuiderwijk et al. (2012), two of the broad categories of open data impediments are “Usability” and “Understand ability”, under which 33 problems were mapped. Under these, we can list at least seven directly related to the lack of capacities from the user side deal with data:

- Data are not understandable for the general public (e.g. related to jargon).
- No explanation of the meaning of data.
- Lack of knowledge about how to interpret the data.
- Lack of skills and capabilities to use the data.
- Lack of statistical knowledge.
- Lack of (domain) knowledge about how to treat the data.

Process (n=instances)	Summary (and example)
Data  Fact <i>Search</i> <i>Browse</i> Extract (n=8)	A dataset is used directly to identify a specific fact of interest. E.g. Finding out the voting history of a local constituency.
Data  Information <i>Manipulate</i> <i>Statistically analyse</i> <i>Visualise</i> <i>Contextualise</i> Report (n=19)	Content from a dataset is given a single representation or interpretation that is reported in text or graphics. E.g. Composing a report that "profile [s] communities of interest within [the local area] as part of the Council's equality & diversity agenda".
Data  Interface <i>Clean, Combine, Subset Data</i> <i>Configure interface tools</i> <i>Write custom code</i> Provide interface (n=26)	An interface is provided allowing interactive representation of a dataset – providing information customized to the user's input. E.g. Creating a searchable interactive online map of stations and former British rail assets.
Data  Data <i>Convert format</i> <i>Filter data</i> <i>Augment/combine data</i> Provide API Dataset for download (n=17)	A derivative dataset is provided for download, or access via an API E.g. I "took Westminster Constituency data, combined it with scraped [General Election] 2005 data from exposed it as RDF."
Data  Service ? Integrate into existing product/service Create new service (n=4)	A service is provided that relies on open data, whilst not necessarily exposing it to the end-user. E.g. Using boundary data from the Census to run an application that forwards reports of Potholes to the correct Highways authority.

www.practicalparticipation.co.uk/kod/report



Figure 3 – Different uses of data, with process, summary and examples. For each type, the number of instances (n) found is detailed. Source: [Davies \(2010\)](#)

- Expert advice is needed to use the data.

[Zuiderwijk e Janssen \(2014\)](#), based on several interviews with government officials, affirm that this lack of capacities in using data may lead to negative effects in open data:

(...) stakeholders do not profit equally from the opening of data. The use of open data is complex, time-intensive and might require certain skills to find, understand and use data. This results in a high threshold for ordinary citizens to make use of the data. Instead journalist and lobbyist have more time and are often skilled in making use of the data. As such open data can be used by certain groups to strengthen their position, instead of creating a level playing field ([ZUIDERWIJK; JANSSEN, 2014](#)).

Inequalities in access to data is starting to raise concerns for those who, for many years, studied the inequalities in access to ICTs. Micheal Gurstein is probably the pioneer in calling attention for this and coining the term *data divide*:

Efforts to extend access to “data” will perhaps inevitably create a “data divide” parallel to the oft-discussed “digital divide” between those who have access to data which could have significance in their daily lives and those who don’t (GURSTEIN, 2011).

A data divide between countries is also mentioned as one of the conclusion of the Open Data Barometer project. Davies, Sharif e Alonso (2015) states that the data divide between countries has grown from the first edition of the evaluation, in 2013, to the second one, in 2014. Countries are clustered into four classes to define their stage in implementing open data policies: High capacity, Emerging and advancing, Capacity constrained and One sided initiative.

Another important publication which shows concerns with data divide is the report by the Data Revolution Group (2014): “There are huge and growing inequalities in access to data and information and in the ability to use it”. The group hosted by the United Nations warns that “Without immediate action, gaps between developed and developing countries, between information-rich and information-poor people, and between the private and public sectors will widen, and risks of harm and abuses of human rights will grow.”

2.9 Linked Data towards Semantic Organization of Open Data

One of the strategies for adding value to data is the interlinking with other datasets. As described by Attard, Orlandi e Auer (2016), Data Interlinking is one of the steps in data cycle that involves value creation. The value creation techniques at this step are Link Discovery, Data Interlinking and Data Integration. “Missing links between data” is also cited as a problem for the use open data. Zuiderwijk et al. (2012) summarized 9 impediments under the category “Linking and combining data”, such as “Data cannot be linked to other data” or “No unique identifiers are available”.

Since the publication of the paper *Linked Data - Design Issues* (BERNERS-LEE, 2006), a new paradigm over online data organization is being pushed: Linked Open Data, better known by its acronym LOD. The main inspiration is exactly the problem of interlinking heterogeneous data over the Web.

Berners-Lee (2006) formulates four basic rules that establishes best practices for linking data:

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.

Thus, a dataset is represented as Linked Open Data if every data unit is identified through dereferenceable HTTP URIs, which should be linked to another in form of a graph. Following the RDF standard, the graph is composed by several connected triples, describing the connection between a subject and an object through a predicate.

The implementation of these rules in several datasets forms an interlinked graph database connected through common elements which is known as LOD Cloud. The last update from the LOD Cloud platform¹¹, in 2014, considers 1014 datasets, using 649 vocabularies as RDF, RDFS, Friend-of-a-friend, Dublin Core and others. Most of the datasets belong to the category Social Web (51%), while Government data represents 18%. The remaining datasets are labelled under Publications (10%), Life sciences (8%), User-generated content (5%), Cross-domain (4%), Media (2%) and Geographic (2%).

Linking data from different datasets over a big virtual cloud is not the only main benefit from the Linked Open Data paradigm. Representing data using shared, linked and standardized metadata can also enable the concept of *Semantic Web*. The idea that computers can understand the meaning of data and documents on the web was already present in the early 2000's, as shown by Berners-Lee, Hendler e Lassila (2001). In this paper, the authors imagine a scenario where several agents present in different devices develop a meaningful communication to solve a problem: setting an appointment with a specialist doctor respecting the agenda and location constraints of several people.

For this scenario to become real, a computer readable definition of how the world works must be developed. This is the main objective of the information science *ontologies*. In the words of Barry Smith, “ontology as a branch of philosophy is the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality”(SMITH, 2003). On the information science field, ontologies came to solve the Tower of Babel problem: different systems with their own concept and relationships definitions wanting to exchange data. According to Smith (2003), “an ontology is a formal theory within which not only definitions but also a supporting framework of axioms is included”. These axioms should explain for computers implicit rules present in the spoken language, e.g., that *a niece is a daughter of a person's brother or sister*.

Currently there are several widely used ontologies, either context specific, as IMDb for movies, or Agrovoc for agriculture, or foundational ontologies as DOLCE or UFO. Though less descriptive and formal than ontologies, several vocabularies are being used to describe semantic content on the Web. Currently, one of the most successful vocabularies is Schema.org, sponsored by Google, Microsoft, Yahoo and Yandex, which claims to be present in over 10 million websites.

On the OGD field, Linked Open Data is still on its first steps. UK's open data portal

¹¹ Available at <http://lod-cloud.net/>.

presents currently 216 datasets in RDF format, which represents 0.93% of all datasets published in data.gov.uk. In his turn, US's data.gov presents 7534 or 3.87% datasets in RDF.

2.10 Conclusion

In this chapter, an overview about Open Data was presented. We highlighted some aspects as definitions and main research problems currently posed. For a more complete view on Open Data, please consult the following selected bibliography:

- *Community Informatics and Open Government Data*, by [Davies e Bawa \(2012\)](#)
- *Open Government Data: The Book*, by [Tauberer \(2014\)](#)
- *A Systematic Review of Open Government Data Initiatives*, by [Attard et al. \(2015\)](#)

This chapter was written basically after a literature research. In the next chapter, we also take conclusions about open data based on the impressions after giving open data classes for social movements activists.

3 Open Data Research Through Data Literacy

The growing tendency of publishing large amounts of data to the Web is so strong that has recently being named as “Data Revolution” ([Data Revolution Group, 2014](#)). Meanwhile, the necessary skills for dealing with data – both from the consuming and publishing sides – are still to be developed by the interested stakeholders. These stakeholders may be government servants or academic researchers, but also members of social movements and civil society, community or grassroots organizations. It is fundamental to guarantee equal opportunities for learning data skills in order to avoid enlarging the data divide, as mentioned in [Section 2.8.2](#).

In the previous chapter, a review about open data was presented, highlighting the main impediments to open data development found in the literature. However, according to the participatory research methodologies ([SCHULER; NAMIOKA, 1993](#); [FALS-BORDA; RAHMAN, 1991](#); [ALVEAR, 2014](#)), involving real users in the research is crucial for understanding the scientific problems and building effective proposals. Thus, we chose to develop a data literacy course in order to get in touch with real open data users, and analyse their motivations, problems and demands regarding open data.

In this chapter, we present the result of a participatory research on open data in form of a data literacy course, as well as theoretical and practical contributions to data literacy. The main contributions are:

- A literature revision about data literacy and related areas;
- An analysis about motivations, impediments and demands from social movements activists regarding open data;
- Theoretical considerations on the application of popular education principles to data literacy; and
- A methodology for researching and teaching open data in the context of social movements.

In the following, we first provide an overview of the Data Literacy field, which is newly being developed. Being a very recent field of academic studies, we propose in [Section 3.2](#) some theoretical contributions, adapting the work of the Brazilian pedagogue Paulo Freire to the Data Literacy field, and defining the concept of *Critical Data Literacy*. In [Section 3.3](#), we present a method for teaching Data Literacy for social movements, which was applied and evaluated. The method includes a research perspective, whose results are shown and discussed in [Section 3.4](#). Finally, conclusion are drawn in [Section 3.5](#).

3.1 An Overview on Data Literacy

The introduction of new digital technologies in the everyday life is an irrefutable reality. Information and communication technologies (ICTs) impact both those who have structure and access to education to enjoy the comfort brought by the ICTs, and those who do not. In order to analyse these impacts from a critical point of view, since the beginning of public Internet in the 1990's, studies about *digital divide* – a term coined to define this social phenomenon – have been developed. This field relied on the concept of *digital inclusion* as a way to overcome the inequalities on access to ICTs¹.

One fundamental step of digital inclusion is *digital literacy*, a term which references a parallel between the act of learning how to read and write – *literacy* – and the act of learning how to use computers. With the growing presence of ICTs in society, specialized questions arise under digital literacy.

From the mid-2000s onwards, governments globally started to publish online big quantities of data (CHIGNARD, 2013). It was the beginning of the worldwide movement towards the so-called open data, understood as the first step of transparency process supporting democratic regimes. As a result of growing need, at the same time, the term *data literacy* started to be coined, even without a formal and widely accepted definition.

The promises brought by open data initiatives relate to a more transparent society, a deeper participative democracy, and possibilities of generating value from data (HUIJ-BOOM; BROEK, 2011). Meanwhile, the severe social inequalities faced all over the world, reflected directly in the education level of the population, creates a strong potential for generating a mass of data illiterates.

Being as data literacy is a new study domain, and thus under construction, there is no established definition for the term. According to the *Data Journalism Handbook*, “data literacy is the ability to consume for knowledge, produce coherently and think critically about data” (GREY; BOUNEGRU; CHAMBERS, 2012). The *Wikipedia* term states that “Data literacy is the ability to read, create and communicate data as information.” Another work highlights the importance of understanding how to produce data (CARLSON et al., 2011).

To the best of our knowledge, the first academic event regarding Data Literacy was the I Data Literacy Workshop, co-located at the 2015 ACM Web Science conference. In one of the published papers, Bhargava e Ignazio (2015) observed that the first mentions of the term *Data Literacy* called the attention for its importance on the context of evaluation of information, together with Information Literacy and Statistical Literacy. In 2004, Schield

¹ There is a vast literature about digital divide, which is out of the scope of this chapter. For a very recent debate on this topic, we recommend Gurstein's paper *Why I'm giving up on the divide* (GURSTEIN, 2015).

reinforced the importance of teaching these three literacies for “students who need to critically evaluate information in arguments” (SCHIELD, 2004).

Wolff, Kortuem e Cavero (2015) describes a data literacy approach applied in schools for young (9–10) and older students. In order to support their narrative and inquiry-based learning approach, a cycle has been developed with the following stages: Problem (define questions), Plan (study/design what to measure), Data (retrieve and clean), Analysis (visualize/look for patterns and Conclusion (interpret/new ideas). After applying the approach to students in the age of 9–10, authors argue that “young learners are capable of working with large data sets” and that data literacy should be included in curriculum of schools.

Vahey, Yarnall e Patton (2006) enforces the difference between statistical and data literacies: while the first one concentrates on applying statistical methods to data, the second one much more to do with the context. These authors also bring the idea of bridging disciplinary divisions with data literacy. A data literacy approach developed in this work starts with students understanding the overall context in social studies, continues with mathematics lessons for formalizing data concepts, and finishes again with social studies to apply the understanding brought by data. Their goals on applying data literacy in the schools is to investigate real problems, formulate and answer data-based questions, use appropriate data, tools and representations, and finally communicate solutions.

A prominent initiative on teaching open data comes from the School of Data, an initiative by Open Knowledge and Peer 2 Peer University. The school works “to empower civil society organizations, journalists and citizens with the skills they need to use data effectively”, under the slogan “Evidence is Power”. In 2014, the School of Data organized 90 events taking place in 30 countries, reaching over 2000 participants. Besides Europe, where most of them happened, School of Data reached places like Lebanon, Nigeria, Indonesia, Mexico, Brazil, Bosnia and Herzegovina, Tanzania and Philippines – training and exploring data about water, elections, and many other issues (School of Data, 2014). Open Knowledge offers courses in Germany, with a special focus on Data Journalism.

Initiatives on open data education have been reported in countries including the United States, the United Kingdom, Spain, Australia, and especially in Denmark, where the focus is on standardization of open data strategies between different government institutions (HUIJBOOM; BROEK, 2011). Fioretti (2011) also notes the importance of using open data in schools, emphasizing that it could help connect school curricula with real life and stimulate active citizenship in the students. The need for some skills to understand data, such as mathematics, was also mentioned. Fioretti proposes two main lines of action: using open data, and producing open data as an official school policy.

3.1.1 Data Literacy and Popular Education

Data literacy initiatives started to be driven since a few years ago, and have been pushed mostly by civil society organization, although there are also governmental efforts. The initial state of this movement is reflected in the academic production, especially when dealing with popular education. The popular education approach for dealing with data literacy is still limited in the available literature.

One exception is a blog post by [Bhargava \(2013\)](#), trying to relate the popular education of Paulo Freire with data literacy. The author introduces the concept of popular data, presenting a synthesis of popular education and its' relationship with appropriation and use of data for decision taking. For him, governments are talking about data, but most of the people are not understanding the conversation. He cites an initiative by the city of Somerville, in Massachusetts, and its ResiStat program, which regularly promotes meetings with the community and stimulates the civic participation via Internet through discussions and data-based decisions. He concludes from this initiative that people can only participate if they have an understanding of tables, graphics and terms related to data. The perspective of popular data, for Bhargava, is oriented by participatory approaches for using data and decision taking that provokes engagement of the population.

Expanding from data to wider ICTs and the relation to popular education, a work by [Adams e Streck \(2010\)](#) affirms the focus of popular education on social transformations through the action-reflection-action of marginalized and oppressed classes. The authors develop their work by questioning the role of ICTs in the production of the current structural conditions, and whether these technologies have the potential for pedagogical mediation seeking the construction of new paradigms. They critically conclude that there are several studies related to education that do not recognize the digital technologies as pedagogical mediations, but as mere tools. According to them, this approach is reductionist, because the pedagogical mediation happens between people through their lived realities, reflecting about it and transforming it. The knowledge production through systematization of experiences and participatory research is emphasized, with a focus on reflection about lived experiences. ICTs, for the authors, “compose a structural reality which conform behaviours, ways of thinking and acting which tends to adapt, modify, recreate and assume emancipatory paradigms”. At the same time, technologies are not neutral and their limits have to be tested, with a constant critical vigilance, and thus popular education cannot but put in the background.

According to [Ferreira e Santos \(2002\)](#), there is a potential for changes in education caused by the wide access to information and knowledge through cyberspace. One of the challenges is to collectively build knowledge between educators and educands, overcoming “bureaucratic separations of authorships between who elaborates, who applies, who clarifies, and who manages the education process”. Authors compare the unidirectional and the

interactive approach in the education field. In the first case, the teacher delivers knowledge and the students have a passive reception role. In the second approach, the complex knowledge network emerged in an educative environment is recognized, and both educators and educands can be authors and co-authors. The concept of co-authorship is recommended to be applied as a praxis to be developed both in on-site and distance education.

3.2 Contributions of Paulo Freire for a Critical Data Literacy²

In the 1960's, in the northeast region of Brazil, the illiteracy rate – percentage of adult people who could not read or write – reached 72.6% (FERRARO; KREIDLOW, 2004). And precisely in that context arose the work of the philosopher Paulo Freire. He characterized the process of literacy education both as technically learning how to read and to write, and as the emancipatory process of understanding and expressing itself in the world: “to learn how to read is to learn how to say the own word. And the own human word imitates the divine word: it creates” (FREIRE, 1987, p.11).

In this section, we aim to trace parallels between the reflections of Freire about literacy education and the critical understanding of the world through data, bringing elements to comprehend the new phenomenon of data literacy. We advise that this is an introductory paper, with a series of limits. The scarce literature about data literacy obliges us to bring inspiration from other sources, and is precisely in this sense that we seek the contributions of alphabet literacy methods to the field of data literacy. The ideas brought here are mostly in the theoretical field. Nevertheless, they came from concrete experiences in teaching open data (TYGEL; CAMPOS; ALVEAR, 2015) and developing information systems for social movements. It should also be noted that Freire's development was driven in a specific context – teaching poor peasants how to read and write, with the intention of raising their consciences – and thus, any adaptation of it for other contexts must take this into account.

3.2.1 Paulo Freire, Literacy and Popular Education

In Latin America, and especially in Brazil, the history of education cannot be told without the name of Paulo Freire. Born in Pernambuco, in 1921, he became worldwide famous for his critical pedagogy, and mostly for the development of the philosophical principles of the Popular Education, the most well known product of which is a literacy method.

The first big experience of the application of the method happened in Angicos, a city in Rio Grande do Norte state in the northeast region of Brazil. In 1963, 300 sugar cane cutters became literate in 45 days, with 40 hours of classes. Subsequently, the then

² This section is adapted from Tygel e Kirsch (2015)

president of Brazil, João Goulart, invited Paulo Freire to organize a National Literacy Plan, with the goal of teaching more than 2 million people to read and write. The plan began in January 1964, but was quickly aborted by the civil-military coup, on the 1st of April 1964. Paulo Freire’s method was substituted by the Brazilian Literacy Method (MOBRAL, in Portuguese), where all the critical view was removed. Paulo Freire was arrested and had to leave the country, returning only in 1980.

In the 1960’s, the traditional literacy method was spread through primers, i.e., booklets containing the content to be taught. This was the central working tool for education, and the focus was on repeating loose words, and in creating decontextualised phrases to reinforce syllables and words. Some classic examples are shown in Table 1.

Table 1 – Decontextualized phrases used in traditional literacy method, in Brazil.

Phrase in Portuguese	Consonant Highlighted	Translation in English
Eva viu a uva	V	Eva saw the grape
O boi baba	B	The ox drool
A ave voa	V	The bird flies

Freire said once that “it is not enough knowing that Eva saw the grape. It is necessary to comprehend what is the position of Eva in the social context, who worked to produce that grape, and who profited from this work” (GADOTTI, 1996). Moreover, Eva is an extremely uncommon name in the northeast region of Brazil, as well as the grape, grown typically in the south of the country. The statement is therefore completely decontextualised, and only encourages the students to memorize it, instead of understanding.

According to Freirean philosophy, the education must be contextualized, i.e., it should arise from the concrete experience of the educands³, and from what is familiar to them. The comprehension of reality does not occur through a mechanical relation between a sign – the written word – and a thing, but by the dialectical interaction subject-reality-subject, where signs and things relate themselves in a political, cultural and economic context. Therefore, the concepts Eva and grape should not be treated abstractly, but inside a context and a reality. In a very simplified way, we can say Freire’s Literacy Method has three stages (SCHUGURENSKY, 2014):

3.2.1.1 Investigation Stage

In this first moment, the themes and words that compose the reality of the educands are defined. These themes must be part of the everyday life of the educands, and be very familiar to them. The primordial idea behind the investigation stage is that the educational process must start from the educands reality. Thus, there is a commitment for educators

³ Some words used in this chapter are specific from Freire’s bibliography: educands (students), educators (teachers), thematisation and problematisation. Debating the origin of them is out of the scope of this work.

to dialogue with educands about themes that have to do with concrete aspects of their lives (CORAZZA, 2003). The generative themes are related to “the universe of speech, culture and place, which must be inquired, surveyed, researched, unveiled” (BRANDÃO, 1985). The research of the vocabulary universe and the identification of keywords of the group or community are the base for developing the generative themes, and thus, for literacy education. They express limit situations, which, for Freire, are mostly oppressive situations (CORAZZA, 2003).

3.2.1.2 Thematisation Stage

This is the stage where the themes are coded and decoded, alongside the discussion about their social meaning in the world. The elaboration of thematic axes relates the generative theme with aspects of a particular or conjunctural reality, and at the same time, organizes the learning process in an articulated sequence. The thematic axes seek to interweave diagnostics and theoretical questions (NUÑEZ, 1998), fostering the dialectic sequence action-reflection-action from the group involved in the learning process. As stated by (FREIRE, 2005), one way of dealing with thematic axes in the learning process is with the coding process, i.e., the representation of the world using symbols as language, drawing or images. Thus, decoding is the process of interpreting these codes. The decoding process generates new information through the production of more abstract higher level coding, based on the knowledge of the world possessed by each educand (BARATO, 1984).

3.2.1.3 Problematisation Stage

In this stage, the focus is on questioning the meanings previously discussed, in a perspective of transformation of the reality. Reflection generates questionings about myths surrounding one owns living reality (FREIRE, 1979). The evinced reality gathered in the Investigation Stage, further coded and decoded, is then understood as something liable to be overcome.

When tackling Paulo Freire’s Literacy Method, the Popular Education perspective must also be mentioned. As a whole educational philosophy, it is inspired in the stages of the literacy method, going deeper in its reflections. In the 1970’s, many experiences of Popular Education in the South Cone – Chile, Argentina, Uruguay and Brazil – generated the reflection of this pedagogy as a permanent process of theorization over the practice in the context of the organization of the popular classes, mainly against dictatorships that were ruling these countries at that time (JARA, 1998). The process of collective construction of knowledge from generative themes and thematic axes, emerged from a lived reality, was named Systematization of Experiences. This should also be included as a fourth stage in the literacy method:

3.2.1.4 Systematisation Stage

In this moment, the lived experience are organized, interpreted and presented, in a communicative sense. Systematizing, more than gathering data and information about a context, is the exercise of theorizing about an experience and deeply analysing it. Systems of thought, information, management and action imposed by dominant powers promote a unique vision of the lived world, and this stage has the aim of elaborating an alternative view (GHISO, 2011). The act of systematizing implies in an evaluation of advances and innovations generated inside a collective experience, which can inspire other groups in other realities. The systematization of experiences presents itself as a method of investigation and “knowledge production, either from local experiences or wider participatory democracy practices, or other forms of political incidence.” (ADAMS; STRECK, 2010).

3.2.2 Parallels between Literacy Education and Data Literacy

After discussing the parallels between both literacies, and the possible contributions of Paulo Freire to the topic, we derive our own definition of Data Literacy in the end of Section 3.2.3.

Before discussing what contributions from Freire can be brought to data literacy, it is necessary to trace some parallels between elements of popular education in general, and Freire’s Literacy Method in particular, and data literacy. In the following, we present three such parallels.

As stated above, literacy education is composed by two complementary and indivisible aspects: the technical ability of reading and writing, and the social emancipatory process of understanding and expressing oneself in the world. In data literacy, we can observe that there are technical capacities related to data manipulation, such as general computer abilities and statistical-mathematical methods, and capacities for critically analysing data, such as understanding the context in which they were generated, and the reality pictured by them.

Looking further into the technical aspect, we can trace another parallel: data literacy entails a higher technological complexity compared with alphabetization. Indeed, a data literacy process can only happen among literate people. While the literacy education process demands only the necessary instruments for reading and writing – a book, a pencil and a paper – the data literacy education normally demands computers, mobile devices, and internet connection. Mathematical reasoning skills are also fundamental to this process. So, we can affirm that data literacy is a technically more complex process than literacy education.

Relating to the absence of literacy, we can say that the social exclusions caused by both kinds of illiteracies have deep differences, as a third parallel. According to the Brazilian

statistical agency, in 2013 8.5% of the population older than 15 years was illiterate. A closer look reveals a high correlation with poverty and regional inequality. In the northeast region, the poorest of the country, the index almost doubles: 16.6%. The rural slice reveals an even higher index: 18.6% of countryside residents are illiterate. Therefore, a correlation between illiteracy, socio-economic standing, and geographical location can be observed.

Finally, concerning both illiteracies, “data illiteracy”, if we can already refer to this term, covers a much larger slice of the population and results in more subtle disadvantages, which however tend to get stronger as far as the open data policies advance. [Gurstein \(2015\)](#) cites two examples where data illiterates were severely affected by the publication of land ownership records as open data, one in Nova Scotia, Canada and another in Bangalore, India. By not having access to data, in both cases, small farmers lost their land to other landowners who checked inconsistencies in the land records and judicially claimed their ownership. The small farmers were elderly and illiterate, and thus also data illiterate. This example meets what affirms [Santos \(2006\)](#), who demystifies the idea that the cyberspace and its informations lie in a decentralized and free access space. For the author, the cyberspace evinces the computer apartheid generated by social inequalities.

3.2.3 A Freirean Inspired Critical Data Literacy

In the following, we present an exercise of adapting key-concepts of Freire’s Literacy Method to what we are going to call critical data literacy. At the end of this section, we derive our own definition for the term. Table 2 shows, in a systematic form, the stages of the literacy method and its possible specializations for data literacy.

Table 2 – Relation between Freire’s Literacy Method and data literacy.

Stage	Literacy	Data Literacy	Result
Investigation	Understanding of educand’s context, and discovery of socially relevant themes in that reality		Survey of vocabulary universe: source for generative themes and thematic axes
Thematisation	Coding and decoding of words and understanding of its social meaning	Coding of the themes into existing (or not) data, and decoding for understanding realities	Generative theme and thematic axis coded as images, film or data
Problematisation	Finding contradictions surrounding the decoded themes, and demystifying the realities	Discovering non-neutrality in data: which aspects are exposed by data, and which are hidden?	Critical view about the themes
Systematisation	Organization, interpreting, and presentation of the lived experience	Organizing and interpreting reality through data, and communicating discoveries	Communication products

3.2.3.1 The Emancipatory Character of Data Literacy

As Freire's method, our data literacy approach has an emancipatory perspective. The literacy concept, as stated above, can be analysed in two dimensions: the technical abilities and the emancipation achieved through the literacy process. Given the high technical complexity of data manipulation, it seems to be a natural tendency that this dimension suppresses the emancipatory one. Immerse in studies involving the use of computers, specialized software, various data sources and statistical methods, there might be a tendency of the educands to leave behind the critical reflection about the social meanings of data in the world, and therefore the emancipatory perspective can be put in background. The emancipatory perspective resulting from data literacy can be materialized in certain abilities acquired by the educands, for example:

Context interpretation: Critical analysis of a specific reality can be more consistently performed based on benchmarking and statistics. As an example, we can cite the topic of land concentration in Brazil. Anyone living rurally in Brazil knows that a few landowners control huge amounts of land. This empirical perception can be better supported if we analyse the agricultural census, which shows that 45% of the arable land is controlled by 1% of landowners, making Brazil one of the countries with the most concentrated land possession in the world.

Questioning of common sense concepts: Many concepts understood as “truth” are build upon data. However, the comprehension about how this data was generated allows a critical eye on these concepts. One example is the concept of Gross Domestic Product (GDP), generally used to distinguish the political importance between countries. Although regarded as the most important measure of a country's economy, it does not consider the income distribution or the environmental consequences of economic development.

Development of new concepts: Through consistent generation of data, it is possible to enlighten invisible realities and establish new concepts. For examples, in 2007, a mapping revealed that almost 2 million people in Brazil worked in self-managed cooperatives, within a solidarity economy context. This data sheds light on other forms of work organization, which normally are hidden or considered small experiments, and allows the establishment of the idea of other possible economic arrangements.

3.2.3.2 Data Literacy Process

Figure 4 shows our proposed critical data literacy process. At the first moment (i), the group observes some context, seeking for elements in common with their reality. Through this view, it is possible to define what kind of data – existing or to be collected – can support and enhance this view. In this moment (ii), data from this context is gathered. The critical analysis of this data (iii) is necessary in order to understand which perspectives

are illuminated by this data, and which are hidden. Finally, after the critical analysis of data, it is possible to look again to the context (iv), see it from another perspective and act towards its transformation. It is important to notice that this is not a linear process, but an iterative one. The last step is always an enhanced realization of the first, and the process should be continued until the objectives are achieved.

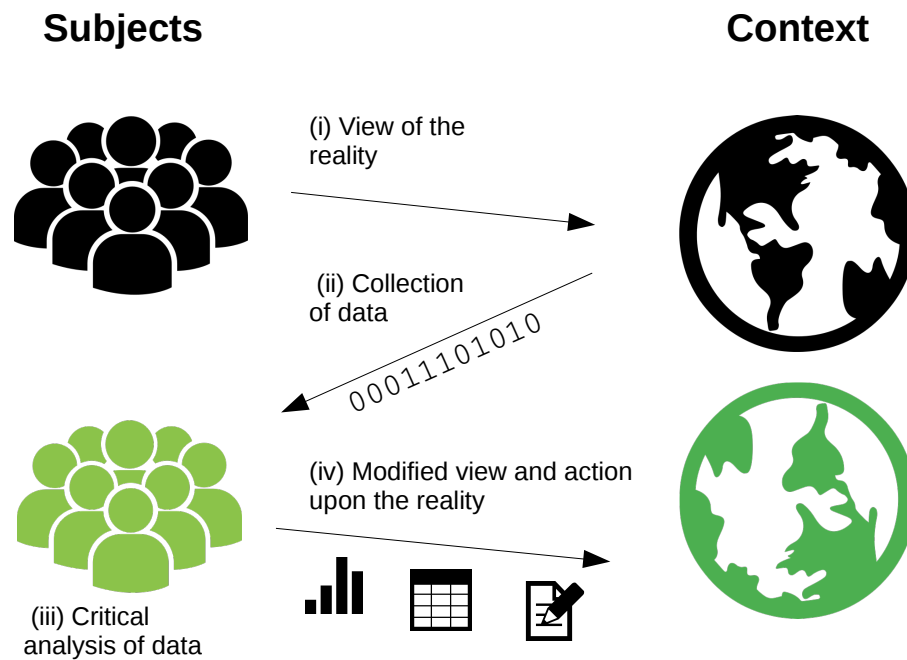


Figure 4 – Critical Data literacy process. Source: [Tygel e Kirsch \(2015\)](#)

3.2.3.3 Data Literacy Stages

Investigation

As already stated, this stage must guarantee that the educational process effectively starts from the educands reality. Just like the grape is not a typical fruit from the northeast region of Brazil, a database is also probably not something that is explicitly part of the everyday life of data educands. (Their personal data, however, are almost definitely registered in one or more databases.) At the same time, it is important to seek in the reality of each educand elements where data can be useful to understand that reality. Considering possible problems in dealing with computers, it is fundamental that the themes to be worked with are of great interest of educands, and have their foundations in daily life. It is also important to find contradictions in this reality that one desires to overcome. Thus, an interesting way of starting this quest is through statistics. For example, as detailed in [Tygel, Campos e Alvear \(2015\)](#), in a data literacy course, the educands

were exposed to statistical informations previously selected about their realities. From this point on, it was shown that, on the one hand, datasets were already part of their life, and on the other hand, that much information known by the educands were omitted by data. Thereby, a data mediated world view is approached, facilitating the most adequate choice of thematic axis to work with.

Thematisation

At this stage, the main goal is to motivate the understanding of the world through data. Either for a local or global reality, about specific or generic themes, data allows an understanding of reality commonly seen as “neutral” or “objective”. At the thematisation stage, it is still possible to keep this aspect, which will be further deconstructed in the problematisation stage.

By elaborating thematic axes, in this stage the aim is to code certain contexts as data and aggregated information, such as statistics, graphics and tables. This coding may lead to more complex decoding about the same theme. A reality can be coded into data, which can be once more coded into aggregated information, and then can be further decoded, generating a modified view over the same reality. It is always important to notice that this process has an intrinsic bias, related to the design choices at data acquisition and processing.

As a result of this stage, it is possible to obtain the generative themes, which in the case of data literacy, are specific context coded into data. This data can be already available as open databases, closed and subject to information access requests, or may also be uncollected data, which could provide some interesting perspectives. The final aim of this stage is to enchant educands with the world of data that represents realities.

Problematisation

After the “enchantment” with the world of data, it is fundamental to problematise it, i.e., to unveil what is behind the scenes when talking about data. In order to use data with critical conscience, it is necessary to know where they came from, how and to what purpose they were generated. Thus, it is possible to politicise the use of data, and deal with them not only from the point of view of a passive user, but from the perspective of someone who is also able to produce data, and with them, “say his word”. The final aim of this stage is to promote a critical view about the chosen theme, understanding the role of data for enlightening certain aspects and hide others. We list here, without any aspiration of completeness, two issues that can serve as a starting point for the problematisation stage:

- *Non-neutrality of Data:* Data are not neutral. The seducing precision and objectivity of data grounded statements almost always hide ideologies and intentions about anything one wants to prove. Thus, it is fundamental to problematise the origin of

data. Are data from the government or from civil society organizations? What was the political position of that organization at the time when data were generated? If it is about scientific data, who funded the research? More complex, but also of great importance, is the knowledge of the methodology used to gather data. Lack of awareness of the methodological approach can lead to misunderstandings and flawed conclusions.

With that information – origin and method – it is possible to infer what was the objective of data generation, where it is not explicit. Producing data is a costly activity, which requires a considerable amount of resources, especially when dealing with big populations and/or wide areas. Therefore, every research that generates data has a very well defined purpose, which must be unveiled and discussed.

Research is designed by specific actors, to reach strategic goals. Similarly, methodologies are designed in order to highlight some aspects, and not others. This is why we can affirm that data resulting from these researches are not neutral, and therefore its non-neutrality must be problematised in a critical perspective of data literacy education.

- *Transparency:* In many cases, the critical use of data will come across the lack of available data. These missing data may not exist, be hidden or poorly organized, which is the case of many governmental data. In order to work critically with data, it is necessary to have conscience of one's rights to access information, which is directly related to transparency policies. Many countries are advancing in this field, publishing their data online and creating laws to guarantee access to information, transparency and open data, with the valuable argument of enhancing democracy and fighting corruption. However, as stated by the Global Open Data Index, only 11% of the assessed datasets in 97 countries are open. Thus, discussing transparency and access to information is a possibility of problematising data literacy.

Systematisation

The systematisation process requires data and information about an experience. In the data literacy context, the ability to put together data retrieved from various external sources with subjective qualitative information empirically obtained should be encouraged.

The systematizing stage should be the conclusion of the whole lived process – investigation, thematisation and problematisation. Of crucial importance is the communication of the results. Data can be exposed in several forms, such as graphics, tables, maps, infographics, music, film or even text. The ability to choose the right way of systematizing and communicating data is certainly a point that should be stressed in data literacy.

3.2.3.4 Definition

Considering the arguments developed in this section, we derive our definition of critical data literacy: Critical Data Literacy is the set of abilities which allows one to use and produce data in a critical way. This set is composed by:

Data reading: *The ability of reading data starts at understanding how data was generated, i.e., which methodologies were used in order to capture data from a context, which facts, measures and dimensions were considered, and at which level of detail, or granularity, data was collected. It also includes understanding who produced it, in which context and why. Data should not be read as objective fact, but as the output of a social process.*

Data processing: *The ability to technically process data is related to the use of computational and statistical tools in order to transform data into information. Linking data with other sources is also an important skill. Data should be processed based on explicit objectives.*

Data communication: *The ability to communicate data comprises finding better matches between data types, such as distributions, temporal series, networks or comparisons, and communications tools, such as text, tables, several types of charts, maps or infographics combining these elements. Communicating data also encompasses a social evaluation of what message should be transmitted to which target audience. Data communication should be done in an ethical, responsible and precise way, in order to avoid misunderstandings or invalid conclusions.*

Data production: *The ability to produce data includes deepening all elements within data reading. Additionally, knowledge about data formats and data publishing tools is required. Generally, data should be published not only respecting the Open Definition, but also offering tools so that non-experts are able to use it.*

3.2.4 Conclusions

The fast spreading of ICTs in the society has, as one of its consequences, a recent publication of massive quantities of data over the Web. These can be either related to governments, through public transparency initiatives, or generated by companies or civil society organizations, or even originated from scientific research. This huge mass of new information brings with it a series of potential benefits, but also major challenges, which are for the most part not as explicit as the benefits. There is an imminent risk of establishing an elite able to profit from these data, interpret it and act in the world through it, while most of the people remain excluded. In this paper, we sought in the work of Paulo Freire inspirations for the construction of a critical data literacy, which incorporates awareness of this challenge.

Future works on this topic includes deriving more tangible examples of the appli-

cation of this methodology in practice, followed by developing a strategy to assess and evaluate the outcomes. From the theoretical point of view, a deep analysis of the digital literacy literature could also bring more elements for data literacy.

It was not by accident that Paulo Freire materialized his Popular Education pedagogy into a literacy method. For him, literacy is not only useful to read words, but to read the world. And imbued precisely by this spirit, we propose an analysis of data literacy based on Freire's Literacy Method. By doing so, we hope to provide a small contribution to the democratization of access to information. Data alone do not change the world, but we believe that people who critically understand the reality through data have better tools to do it.

3.3 Teaching Open Data for Social Movements: action and research for open data engagement⁴

Motivated by research on use and publication of open data by social movements and grounded on popular education principles, an open data course was developed. According to the dialogicity principle, the course objective is double: (i) to tackle the issue of open data education, indicated to be one of the factors hindering the use of open data; and (ii) to use the time in training to observe the activists using data and gather information for the research.

The course programme was elaborated seeking a balance between the social aspects of the use of data, the principal motivation, and the technical issues that are inherent in the tools for data manipulation. The methodology switches between expository stages and individual and collective activities by the students. It is expected that the students can at least achieve a critical view about data, understand the possibilities and limits of its use, be aware of the political questions involved in data production and publishing, and, finally, have a technical starting point for manipulating data.

The course is divided into four stages of four hours each, but can be adjusted to needs of the people involved. A website containing teaching materials, links to data sources, and a discussion forum was developed, which in each presentation of the course is supplemented with more data. Only two requirements are asked of people interested in attending the course: a basic knowledge of informatics (web navigation) and access to a computer (which could also be offered by the organizers). Good quality internet access provided by the organizers is also highly desirable.

⁴ This section is adapted from [Tygel, Campos e Alvear \(2015\)](#)

3.3.1 First Stage – Introduction

The first stage starts with a short description of the course, and the participants are informed that they will also be contributing to a research project. This stage is intended to get people on the same level, by discussing the sociotechnical and political aspects of data. The aim is to start from the educands' own experience, as suggested by the Popular Education approach. For this, all participants are asked to present themselves, state their expectations and why he or she decided to take part.

Afterwards, a challenge is proposed: some socially relevant statistical results are presented (see Listing 3), and the educands are asked to find the data sources related to those figures. Following the inverse path (information to data, rather than the opposite), we expect to raise curiosity and show, in practice, the importance of knowing what is behind the statistics.

Table 3 – Examples of data driven statements used to stimulate a critical view of data sources (based on Brazilian statistics agencies)

1	0.9% of the biggest landowners own 45% of arable land in Brazil
2	In Brazil, white men earn more than white women, who earn more than black men, who earn more than black women
3	77% of young people killed in 2011 in Brazil were black
4	46.7% of Brazilian exportation in 2013 were basic products, 12.6% were semi- manufactured, and 38.4% were manufactured

In the sequel, several open data related topics are discussed:

How does data arise: a data path is presented, from the occurrence of something, passing through its systematization to its publication. Concepts such as facts, dimensions, and measures are discussed, together with the political motivations and consequences of those design choices. This topic is intended to put data neutrality in question, by showing that data produced by research is an outcome of several choices, made according to some goal.

Data visualization: the same dataset can be observed in many ways, and the conclusions to which one may come heavily depend on this. Visualizing data as tables, graphics, networks (graphs), or maps may reveal different aspects and induce several kinds of conclusions.

Open Data: In this topic, we motivate the understanding of open data using analogies (see Table 4). In the sequel, we define open data according to the David Eaves' three rules: data must be findable in the Web, published in machine readable formats, and cannot have licenses which prevent re-use (Eaves, 2009). A debate about linking and semantically

Table 4 – Open and closed analogies to help understand what open data is.

Open	Closed
Text in digital format (txt, odt, doc)	Printed Text
Presentation in editable format (odp, ppt)	Presentation in PDF format
Source code	Executable software
Raw Data	Information (statistics, graphics, maps)

marking data through the use of Linked Open Data (LOD) is also proposed with examples. Transparency Policy: At this point, we present the context of open data in Brazil and in the world, especially through transparency policies. It starts with the Freedom of Information Act (FoIA), and goes up to Internet governance, with the recent Brazilian regulation¹ based on three foundations: net neutrality, privacy and freedom of expression. International efforts on transparency, such as the Open Government Partnership (OGP) are also presented.

Synthesis: After presenting all topics, educands are asked to discuss how open data is related to their activism.

3.3.2 Second Stage – Data Sources

The second stage of the course is dedicated to an overview of some important datasets on the Internet. It is worth noting that some of them are not “open” by the classical definition (Eaves, 2009), mainly because the raw data is not available for download. However, when an aggregate data querying system is offered, it makes data even more useful for common user than if raw data was available.

Different forms of accessing data are discussed. We recognize that, in respect to data access means, there is a trade-off between the ease of analysing data and the autonomy one can have in assessing one’s own conclusions. When a database is published as raw data, following all open data principles, this still might not help a citizen who wants to know how much was spent on education in his city. Large volumes of data coded in specialized formats (e.g. R, SPSS, SAS, SQL, XML, RDF) allow a high level of autonomy in the analysis, but special skills are needed to work with it. On the contrary, aggregate data, reports and charts allow people to have access to this information, but it has already passed through someone else’s filter. Figure 1 depicts this debate.

Besides the means of data access, we propose a classification of data according to the type of provider: Data produced by the state: This is the wider category, since the state has structural conditions and legal liability to produce data. In Brazil, the biggest data producer is the Brazilian Institute of Geography and Statistics (IBGE, in Portuguese),

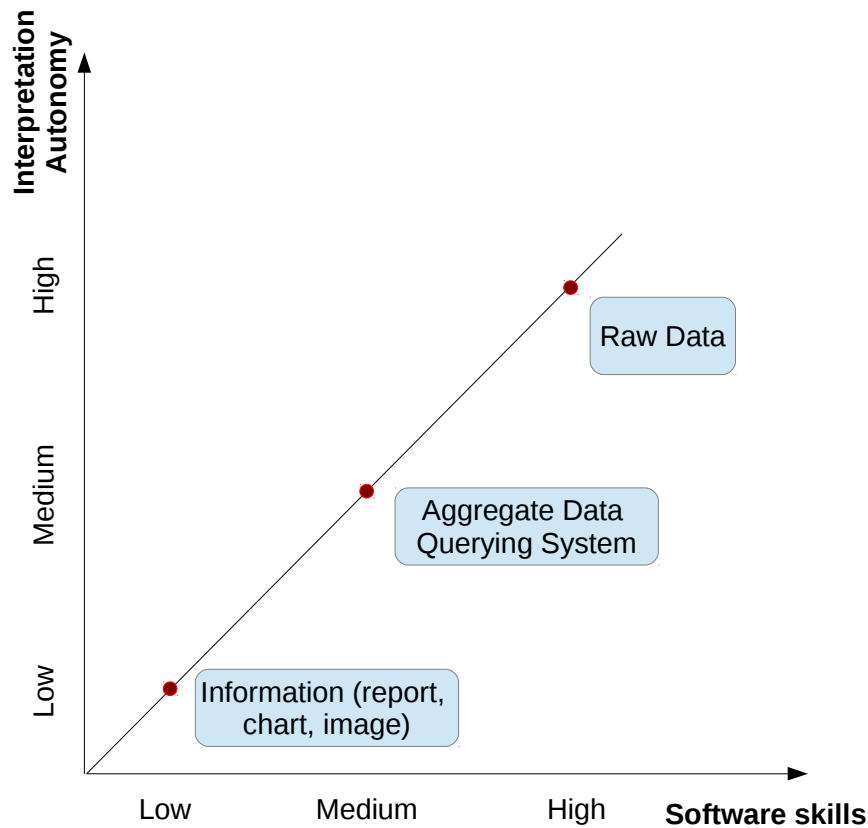


Figure 5 – Trade-off between interpretation autonomy and software skills needed.

responsible for demographic, economic, geographic and many other sorts of data. The Unified Health System (SUS, in Portuguese) is also an important data generator, mostly about health and illnesses. Worldwide, the United Nations (UN) and the World Bank are also important data suppliers. Even though they are not governments, most of their data is compiled from country data. It is important to emphasize that this kind of data carries with it the visions and ideologies of those who generated it. All the design choices made during the data production, including definition of facts, dimensions, and measures, in some degree follows the government intentions.

Data produced by the state and shown by society: In many cases, data produced by the state is not open, and when it is open, there are no tools for the citizens to easily analyse and take their own conclusions. Specialists are needed in order to translate data into useful information. In order to tackle this issue, many society-driven applications using official data have been recently released. In many cases, they help visualize data in a way that leads to conclusions against the states' interest. One example is the Brazilian's "Congress Owners" application. Based on raw (and hard to analyse) donation data published by

Table 5 – Examples of society driven databases, used by social movements with several purposes.

Initiative	Publisher
Environmental Injustice Maps (Brazil)	Fiocruz and FASE
Environmental Conflicts Maps (Minas Gerais, BR)	Federal University of Minas Gerais, Brazil (in partnership with a number of social movements)
Atlas of Environmental Justice	23 worldwide organizations (see http://www.ejolt.org/section/team for a complete list)
Agroecology initiatives	ANA/ABA (Brazilian national social movements related to agroecology)
Land Conflicts	Comissão Pastoral da Terra (CPT)

Electoral Justice, a civil society organization has developed an application where people can easily access and visualize the amount of donations received by politicians and parties, or paid by enterprises or economic sector.

Data produced by society: The case where organized groups of the civil society produce their own data is interesting because: (i) as in the case of state data, data produced by the civil society contains its ideological influences in the design choices; (ii) it allows other perspectives on subjects already explained by the state.

Data related stories can oppose well established hegemonic opinions. One example is the Brazilian Map of Environmental Injustice. Agribusiness is considered to be a good development alternative for the country, based on its relevant contribution to the gross domestic product. The map shows 82 occurrences of Environmental Injustice related to the agribusiness (from a total of 501), where activities of this sector cause damage to poor communities and/or to the natural habitat. Table 5 shows a number of society driven databases. It is worth noting that, in some cases, the funding for building those databases comes from the Government. In principle, we consider that this does not hurt society's autonomy and freedom to put their views forward in the design process.

In the final activity of this stage, educands are asked to add new data sources to the course web page, according to their interests. New sources can come from students'

experiences, or be searched for during the class time. However, it is important to find the exact link, since this is reported to be a difficulty, as it will be seen later.

3.3.3 Third Stage – Tools

In the third stage, the focus is on tools for manipulating data. The goal is to present the means to work with the raw or aggregate data resulting from queries. It begins with an introduction to the Comma Separated Values (CSV) format, which is an open, universal and easy-to-use way of exchanging tabular data. Concepts such as primary and foreign keys are also discussed, in order to help comprehend how relationship between tables and databases can be made. Nevertheless, database design is beyond our scope.

This is an essentially practical stage. Several tools are presented, so that each student can choose which one he or she wants to work with, according to individual interests and ability with computers.

The first tool presented is a spreadsheet editor. The task consists in downloading a CSV sheet with a two dimensional table (production of food in tons, by type of food and year) and drawing a line chart. Students are also asked to plot percentage changes between first and last year production. The second part of the task consists in working with dynamic tables, which allows building analysis frameworks with more than two dimensions.

Other tools presented are related to map building and infographics drawing. Sometimes a mathematical background revision is necessary, since working with number variations requires some previous knowledge of percentages.

3.3.4 Fourth Stage – Final Work

The fourth and final stage is dedicated to a jointly decided activity. The goal is to develop some data based communication product, based on the three previous stages. Ideally, there should be more than one facilitator in the room, so that the work can be divided into groups, with each group being accompanied by one instructor.

Suggested options include: writing news text based on data, and building infographics and maps on specific subjects. The intentionality – what and why we want to communicate – is discussed first. Then, we evaluate the feasibility of the task – is there data about this subject? – and finally, the communication instrument is chosen. In the end, results are presented and an evaluation of the course is done.

The next section brings an analysis of presentations of this course, and draws out some research results based on the experiences gained.

Table 6 – Summary of the presentations of the open data course for social movements.

N.	Kind of Place	City	Duration (h) and time distribution	Participants enrolled	Forms responded
1	Union	Rio de Janeiro	16h (four days at night)	6	2
2	University	Rio de Janeiro	16h (two full days)	11	3
3	University	Vitória	16h (four days at night)	13	4
4	University	Porto Alegre	12h (one half day/one full day)	12	3
5	Union	Rio de Janeiro	8h (two days at night)	10	3
Total			68 h	52	15

3.4 Open Data Clues from the Field⁵

In this section, we describe the application and the systematized results of the above detailed open data course.

The course was presented five times in the second semester of 2014, in Brazil. While three presentations happened in Rio de Janeiro, one was held in Vitória (state of Espírito Santo) and another in Porto Alegre (state of Rio Grande do Sul). Two presentations were held in a workers union and three in universities, organized by groups who work with social movements in extension projects. A total of 52 students enrolled and participated in at least one stage. There were no fees to pay, and the only requirements were basic informatics knowledge and access to a computer, sometimes provided by the organizers. Table 6 shows a summary of the presentations.

The analysis will be based on two instruments: an evaluation questionnaire that all students were asked to fill in, and a participant observation gathered during the presentations. The goal of the analysis is to respond to the research questions: (i) why social movements use data (motivations); (ii) what are the mains problems (impediments); and (iii) what could be done to enhance the use (improvements). Also, the evaluations about the course can be used to improve it.

⁵ This section is adapted from Tygel, Campos e Alvear (2015)

Table 7 – Questionnaire answered by course attendants. All the numerical results are in over a base of 15 ($n = 15$), and N/A means “not applicable”.

#	Question	Mean (maximum - minimum)
1	Age	31 (25–48)
2	Knowledge of informatics (1: poor knowledge – 5: good knowledge)	2.7 (1–5)
3	Work/Profession/Activism	N/A
4	Why have you attended to the course? Why do you think open data is important?	N/A
5	Educator’s performance (didactics, material, knowledge, punctuality) (1: poor – 5: very good)	4.5 (3–5)
6	Self performance (participation, attention, punctuality) (1: poor – 5: very good)	3.3 (1–4)
7	Was the subject according to your expectations? (1: totally distinct – 5: totally according)	4.6 (2–5)
8	What is the main impediment perceived by using data?	N/A
9	How do you imagine that the use of data could be improved?	N/A

3.4.1 Questionnaire based analysis

All the participants were requested to answer a questionnaire after attending the course. Thus, we assume that the opinions given are strongly influenced by the discussions held over the course. This decision was taken having in mind that: (i) open data is not a subject of the educands’ everyday life; so, answering before the course could lead to meaningless results; (ii) according to the popular education methods, we expect each educand to be able to relate content unseen before with their experiences, and in the end to synthesize their own conclusions about the process. Table 7 shows the questionnaire and the mean, maximum, and minimum values for numerical questions.

The median age of participants was 31 years, with the youngest being 25 years old and the oldest 48. They considered themselves to have medium knowledge of informatics. Before enrolling, participants were asked to have some informatics knowledge, but no admission tests were given.

Some participants were exclusively activists or academics, but most of them were activists with some academic involvement. There were journalists, lawyers and social

scientists, all engaged with some social movement. No participant had formal informatics training, meaning that no one was an informatics expert.

The teacher's performance was well rated, but this was somehow expected in a free course. On the other hand, no one rated him or herself with very good participation performance. In Question 7, only one participant seemed to have very different expectations about the course content. All the others marked 4 or 5, indicating that open data is not so distant from non-informatics people's lives, at least for those who answered the questionnaire.

In order to analyse questions 4, 8 and 9, we will pick answer elements and classify them according to research goals: motivations, impediments, and improvements. Question 4 was aimed to catch motivations, but impediments and improvements were also cited. Question 8 raised only impediments, and Question 9 only improvements, as intended. An effort was made to extract concrete elements from the discursive text. An equilibrium was sought between merging similar statements and not losing the diversity of opinion. These concrete elements extracted can be seen in the Appendix, in Tables 13, 14, and 15.

Sometimes, the separation between the classes is not very clear. All impediments (e.g. "Open Data Portal is hard to use") have implicit improvements (e.g. "Open Data Portal could improve usability"), as all improvements also have implicit impediments. Some motivations (e.g. "Use spending data to fight corruption") also could be interpreted as impediments (e.g. "Few spending data is available") or improvements (e.g. "More spending data must be made available"). We tried to classify according to the respondent's intention.

3.4.2 Observation based analysis

In this section, some remarks are made based on the 68 hours observation of the course. This observation was driven inspired by the ethnographic method of participant observation (ATKINSON; HAMMERSLEY, 1994). Within this approach, the researcher plays an established participant role in the studied scene, in this case, as an educator, taking field notes during the class. Ethnography inspired methods are complementary to objective and quantitative evaluations since, according to Atkinson e Hammersley (1994), ethnography deals with the "analysis of data that involves explicit interpretation of the meanings and functions of human actions", and "represents a uniquely humanistic, interpretive approach, as opposed to supposedly 'scientific' and 'positivist' positions." Since two of our research questions deal with human actions and feelings – what are the motivations of social movements for using open data and what are impediments that block a wider and better use – we considered the participant observation an appropriate methodological direction. We aimed to comprehend the point of view of the educands, and this was done from the educator stance, which certainly influenced the analysis.

As described earlier, in the first stages of the course participants are shown statistical statements (see Table 3) and are asked to search for data that generated those figures. Below, we list some behaviours observed:

- The first impulse of users is to paste the phrases directly into a web search engine. Normally, the results are news commenting that statement, or reports containing that information, and never the actual data source.
- For some people, it is difficult to understand the difference between the statements and the data sources from which they were originated. One way to overcome this misunderstanding is to slightly rephrase the statement and ask what would be the new figures. For example, relating to statement 1 (Listing 1), we would ask: “how much land do the 0.1% of the biggest landowners possess?”.
- Overall, only few people reached the actual data source. This shows that one of the main problems of data sets and their query/download systems is that they are frequently hidden in the deep web, i.e., regular search machines cannot find them.

In the second stage of the course, some data sources are presented and divided into three categories. About this stage, we would like to remark:

- In general, although interested, users are unfamiliar or unaware of data sources. This ignorance is, as expected, worse for society driven OGD based applications, and for data produced by social movements, which usually have no official means of dissemination;
- Students were stimulated to add new data sources to the course website, according to their own interests or activism. In some cases, participants inserted already known data sources, but in most cases data sources were found during the activity.

The third practical stage revealed one of the strongest difficulties in open data usage: the manipulation of software tools, particularly of spreadsheets. The knowledge about CSV tabular files, considered as a fundamental skill to use data on different systems, was practically absent. This problem got even worse because of the inability of the most-used proprietary spreadsheet application (MS Excel) to deal with such kind of file. LibreOffice, its open source counterpart, facilitates this task.

Another issue that was highlighted at this stage was the mathematical difficulty faced by most of the students. Dealing with statistical open data requires, most of the time, simple mathematical operations. Therefore, sometimes a small revision of percentage was necessary.

Unfortunately, the fourth stage of the course did not work as expected. This stage was only reached in two of the five presentations described in Table 6. In the first one, students, mainly journalists, decided to individually write stories and impressions about open data. They were published in the course website. The second experience reached

closer to the goal: participants decided to investigate a local case of environmental injustice. Data about enterprises, population health, environmental licensing and other issues were gathered, but no final product was obtained. In the remaining three presentations, the time ran over twice, and once the students said they were tired, as this course was run on two full days, at a weekend.

One possible way to overcome this issue is to propose this work at the beginning of the course and organize tasks during the other stages. This has the advantage of motivating students with a concrete problem during the course. Nevertheless, the challenge remains: how to prepare the course without predefining the problem. Another option would be to increase the number of hours, which would depend on participants' availability.

3.4.3 Synthesis

As explained above, by a simple rephrasing, an impediment or a motivation can turn into an improvement. By doing a careful analysis of Tables 13, 14, and 15 (see the Appendix), an improvement classification tree was built. It is aimed at orienting actions for the engagement of social movements in open data in the Brazilian context. The classification tree can be seen in Figure 6.

The IT Specific issues are divided into Training and Open Government Data (OGD) Publication. The first class encompasses cited impediments which could be approached with educational investments, and the second is related to actions to be taken by government data publishers. Our proposed course tackles all cited educational demands, except data publishing, since it still demands a higher level of informatics knowledge. As to OGD Publication related issues, the need for better search engines was the most cited enhancement.

The right side of the tree presents general issues related to Transparency Policies and Open Data Publicity. We can conclude that in order to improve open data usage, actions must be taken far above data level. In this case, the whole structure for information access must be enhanced. Difficulties in claiming the FoIA within local government levels were reported, as well as accessing information on private foundations that run on public money. Finally, many participants suggested that more publicity on open data already available would also improve usage.

Some improvements related to OGD publication could be addressed by using new technologies being developed under the Linked Open Data (LOD) framework. By semantically annotating data with commonly used vocabularies and ontologies, the LOD approach offers the technical means to link different data sources and jointly query them. A solid set of tools to implement LOD is being developed (AUER; BRYL; TRAMP, 2014), but strong efforts must be made to hide the complexity of the representation and to highlight

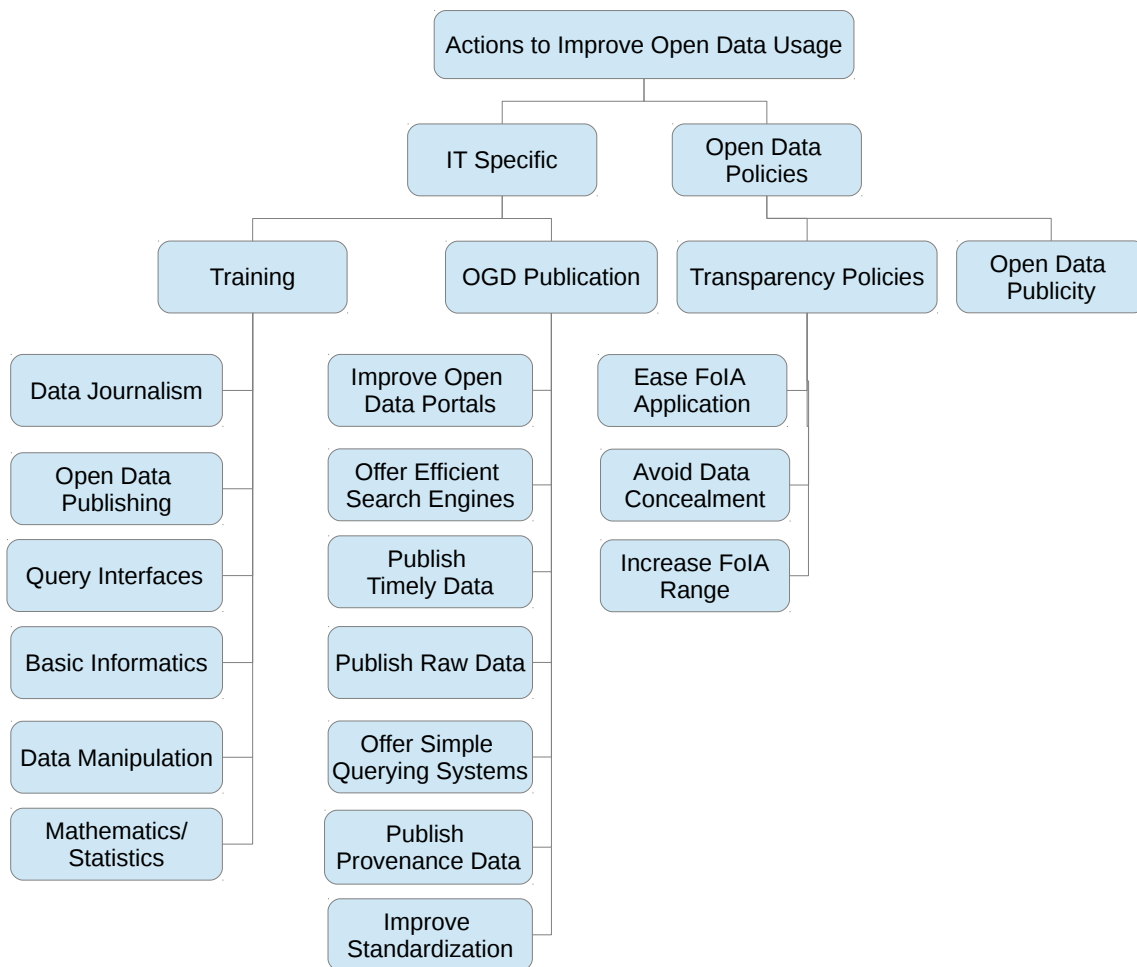


Figure 6 – Classification tree for open data engagement actions, systematized from Tables 6, 7 and 8. The first classification is a distinction between Information Technologies (IT) Specific and Open Data Policies related issues. There is no intention to imply a duality between social and technical issues, however, one can easily recognize that some elements are directly related to information technology, and others are not.

its benefits, so that it can be recognised as a viable option. Other improvements are only possible through the effective political willingness of governments to be transparent.

As a methodological approach for research in informatics, the course was found to be an efficient tool, since it accomplished its dialogical function indicated by the Popular Education theory. At the same time as they were subjects on an open data education action, the educands that participated on the course acted as objects in a research project. On one side the dialogical approach resulted in a set of appointments for open data publishers; on the other side, in a satisfactory educational experience, as shown by the

good educator evaluation (Table 7) and by the rich answers collected (Tables 13, 14, and 15).

3.5 Conclusion

In this chapter, we presented both a participatory research on open data and theoretical and practical contributions to data literacy. Together with the open data landscape presented in the previous chapter, we formed a solid view over the question dealt in this thesis. Considering that open data organization is an important question that hinders the achievement of open data promises, in the next chapters we will present a related approach. The following chapter presents a literature review on methods for dealing with semantic metadata, and is followed by Chapter 5, where we present the Semantic Tags for Open Data – STODaP – approach.

4 Semantic Metadata for Open Data Description

In the previous chapter, the issue of building open data skills was tackled through the development of a data literacy course as part of a participatory research. One of the results of this research points out a significant problem related to the organization of ODPs. Following this motivation, we divide this chapter in two parts, covering:

- A literature investigation over the possible strategies to deal with semantic open data organization;
- An analysis of the current status of metadata in ODPs.

Particularly, in the first section we selected works related to semantic enrichment of metadata in ODPs, in order to position the main contribution of this thesis presented in the following chapter. The section starts with some preliminary discussions regarding semantics and metadata. Then, a characterization of our contribution is driven, in order to delimit the related research topics. After this characterization, we present in each section one processing step, highlighting the main related works, their gaps and relations to this work. We start with Assessment of Metadata in [Subsection 4.1.2](#), followed by Metadata Cleanup in [Subsection 4.1.3](#), Metadata Reconciliation in [Subsection 4.1.4](#) and finally with Structure Emergence in [Subsection 4.1.5](#).

The second section aims to bring light over the current state of metadata use in Open Data Portals. Based on the CKAN Census, we profile 87 portals and analyse several aspects regarding metadata. The analysis embraces not only local aspects, such as use, reuse and similarity within a portal ([Subsection 4.2.1](#)), but also global features between portals, such as coincident metadata and expressiveness ([Subsection 4.2.2](#)).

4.1 A Literature Review on Semantic Metadata

It is unnecessary to argue that good metadata are crucial for making data usable. By *good* we can give as example a series of quality attributes as clean, well organized, detailed, complete, accessible, and meaningful. Intuitively, metadata is meaningful if it brings new information – meaning – for data. If a consumer asks: “Which banana do you have?”, and the seller answers: “The yellow one!”, this is barely meaningful, since almost all types of banana are yellow. However, if the seller answers: “I have *Cavendish*, *Gros Michel*, *Latacan*, and *Cambuta*, which one do you prefer?”, there is much more information accessible through the types of bananas, including colour, size, countries origin, among others.

On the Web context, the way of enhancing the meaning of an object is to connect it to the Semantic Web, through the Linked Open Data Cloud, as detailed in Section 2.9. This procedure is also called Semantic Enrichment or Semantic Lifting. [Limpens, Gandon e Buffa \(2013\)](#) state a series of motivations for semantically enriching tags in the context of folksonomies, considering data generators, data curators and end-users:

1. enriching tag-based search results with spelling variants and hyponyms, or
2. suggesting related tags to extend the search, or
3. semantically organizing tags to guide novice users in a given domain more efficiently than with flat lists of tags or occurrence-based tag clouds, or
4. assisting disambiguation.

A more detailed view about problems caused by the absence of semantics in metadata is described by [Marchetti e Rosella \(2007\)](#). According to them, there are six categories of problems:

1. **Polysemy:** the same word can refer to different concepts (the word 'field' can refer to a piece of land cleared of trees and usually enclosed, but also to a branch of knowledge);
2. **Synonymy:** the same concept can be pointed out using different words ('auto', 'car', 'machine' are three different words that refer to the same concept: a four wheels vehicle);
3. **Different lexical forms:** the same concept can be referred to by different noun forms, for instance plural nouns ('car'/'cars'), different verb conjugation ('buy'/'buying'), name-adjective couple ('energy'/'energetic'), multiple words ('pc'/'personal computer') and so on;
4. **Misspelling errors or alternate spellings:** typing errors that occurs when we write a word ('staton' in place of 'station') or different possible spelling of the same word ('color'/'colour');
5. **Different levels of precision:** the specificity of the word chosen to tag a resource ('jazz' is more specific than 'music');
6. **Different kinds of tag-to-resource association:** implicit kinds of relations that links a tag to a specific resource ('interesting' expresses an opinion on the resource, 'car' expresses the topic of the resource and so on) ([MARCHETTI; ROSELLA, 2007](#)).

Around ten years ago, the discussion about semantifying folksonomies started. Probably one of the most important work at that time was *Ontology of Folksonomy: a mash-up of Apples and Oranges* ([GRUBBER, 2007](#)). This work, published first on the web in November 2005, aimed to clear up a false contradiction between ontologies, as the enabling technology for sharing information on the Semantic Web, and Folksonomies, a typical phenomenon of the Social Web representing data emerged from shared information. It is perfectly reasonable that these two concepts could be understood as a contradiction: while ontologies are formally built by domain experts and ontology engineers, folksonomies are freely constructed by users. After clarifying the role of each concept, Grubber defines the ontology of folksonomy, whose central element is *Tagging*, which is an activity involving an object O , an user U , a tag T and a system S . The possibility of qualifying a tagging is also mentioned, for example, with a negative tagging.

Another important work introducing this topic is "Ontologies are us: A unified model of social networks and semantics" ([MIKA, 2007](#)), also published first in November

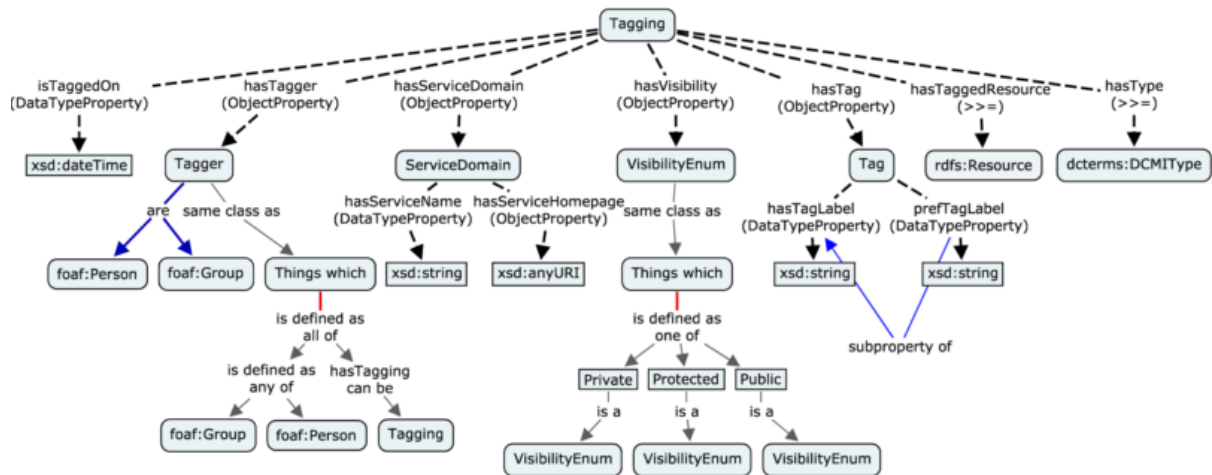


Figure 7 – Tagging Ontology (KNERR, 2006)

2005. Mika also disagrees that ontologies and folksonomies are contradictory, but differently from Grubber, for who both are distinct concepts (Apples and Orange) that can be united, he states that “folksonomies are ontologies”. In order to justify it, the author cites a set of broad ontology definitions, and classifies folksonomies in these definitions as “lightweight, dynamic and limited in sharing scope”.

In the sequence of these papers, several authors tried to define tagging ontologies. Wu, Zhang e Yu (2006) added a time dimension to the tagging model. And to the best of our knowledge, Newman (2005) was the first work to propose an ontology for tagging. This work was further extended by Knerr (2006), who proposed the Tagging Ontology, depicted in Figure 7. All the dimensions proposed by Wu, Zhang e Yu (2006) and Grubber (2007) are present and further detailed in the ontology.

Although crucial, models are not enough to enable the creation of ontologies emerged from collaborative tagging. Halpin, Robu e Shepherd (2007) analysed the dynamics of collaborative tagging, in order to determine the possibilities of extracting knowledge.

It is important to notice that until this point, tagging ontologies were concerned with organizing the knowledge contained in the tagging activity. The MOAT architecture was the first to explicitly include the concept of tag meaning, associating each tagging element to a LOD resource (PASSANT, 2008).

A review about semantic tagging initiatives by Kim et al. (2008) compared the different types and relations proposed by the works until 2008, and was updated by Kim et al. (2011).

The most recent attempt to build a tag ontology is the Modular Unified Tagging Ontology (MUTO)¹, shown in Figure 8, and described by Lohmann, Díaz e Aedo (2011).

¹ <<http://muto.socialtagging.org/core/v1.html>>

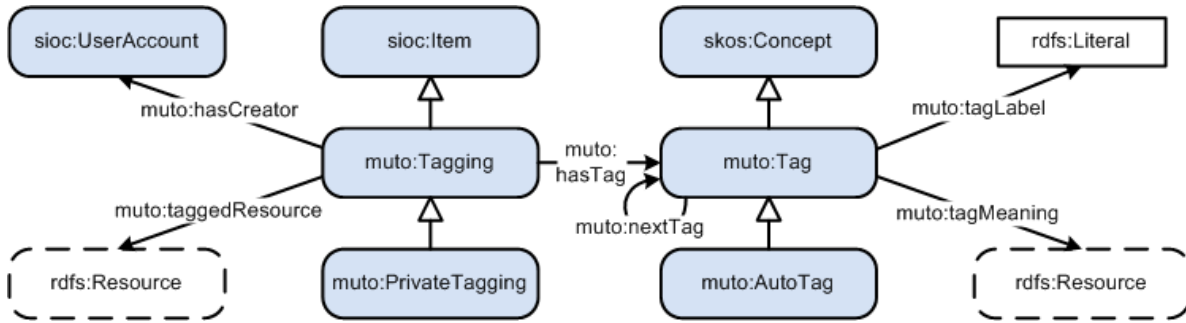


Figure 8 – MUTO Ontology

It incorporates suggestions of several previous models into a unified model, and is strongly based on wide used ontologies, such as Dublin Core, SKOS and SIOC.

Finally, in the context of tagging semantics, it is also important to discuss the nature of the relation between tags and tagged resources. To the best of our knowledge, none of the proposed tagging ontologies incorporates the possibility of qualifying this relationship. [Marchetti e Rosella \(2007\)](#) points out the question of “implicit kinds of relations that links a tag to a specific resource” with an example: “*interesting* expresses an opinion on the resource, *car* expresses the topic of the resource.”

4.1.1 Characterization of the Contribution

The main contribution of this thesis is a semantic approach for organizing metadata in Open Data Portals, specifically regarding tags. Thus, the following topics are considered to be related works, and will be analysed in the following:

- Metadata assessment;
- Metadata Clean Up: spell-checking, detection of similar words, special characters equalization;
- Metadata Reconciliation: synonym detection, automatic translation, vocabulary connection;
- Structure emergence: finding relation between tags.

Regarding the related bibliography, it is necessary to highlight that the vast majority of scientific works about tagging and semantics focus on a different kind of context in relation to ours. In this folksonomy case, tags are attributed to resources by the crowd, passing through a crowd-selection mechanism, which can enhance the tagging quality, but inserts some inherent noise. This is applicable to platforms such as *del.icio.us* or *flickr*, where several users can tag the same resource. However, in the open data portals context, tags are only attributed by system managers. Although less noisy, this procedure is biased by few taggers.

4.1.2 Metadata Assessment

An important step in working with metadata is to develop methods for evaluating quality aspects of it. [Reiche e Hofig \(2013\)](#) implemented quality metrics for metadata in ODPs which can be assessed automatically. In this work, authors measured completeness, weighted completeness, accuracy, richness of information and accessibility as defined by [Ochoa e Duval \(2006\)](#). Although the metrics definition are significant, their implementation in an automatic context is simplified, which in practice turns the accuracy and richness of information, for example, very weak.

In relation to the metrics for tagging environments, some related ideas could be found in the literature. For example, [Umbrich, Neumaier e Polleres \(2015\)](#) present a framework to evaluate the quality of ODPs. Among the applied quality metrics, three of them – *Usage*, *Completeness* and *Accuracy* – are related to metadata keys, which tags are part of. *Usage* establishes which metadata keys are actually used in a portal; *Completeness* evaluates the presence of non empty values; and *Accuracy* checks if metadata adequately describes the data. However, this metric is not applied for tags.

Laniado and Mika did a similar analysis over hashtags on Twitter ([LANIADO; MIKA, 2010](#)). Their work is focused in answering if Twitter hashtags constitute *strong identifiers* for the semantic web. To achieve this, four metrics are used: frequency of hashtags; specificity, which is the deviation from the use of them without being a hashtag; consistency; and stability over time.

[Colpaert et al. \(2014\)](#) presented a method for calculating interoperability between ODPs based in the identifiers used in datasets. The metric developed by the authors take into account if the same identifiers were used to represent the same concepts in a dataset, and then calculates an interoperability metric.

These works are taken into account in [Section 4.2](#), where we derive an extensive analysis over ODP metadata.

4.1.3 Metadata Clean-up

When dealing with metadata of large datasets, a cleaning up procedure is always necessary. There are several strategies for cleaning up tags described in the literature.

[Angeletou \(2008\)](#) describes a Lexical Processing procedure to clean up tags containing special characters, numbers, concatenated tags or tags with spaces. Two steps are proposed in this work: the first is called Lexical Isolation, which uses a set of heuristics to determine if tags have a potential to become semantic identifiers. The following step is called Lexical Normalisation, which aims to produce a list of possible lexical representations for each tag, considering plural and singular forms, different verb tenses, and others.

Although the focus of [Specia et al. \(2007\)](#) lies on creating tags clusters, the procedure includes a pre-processing phase. As the previous work, the first step consists in removing tags with low chances of being mapped in an on ontology. In the sequence, a series of heuristics are used to group morphologically very similar tags, including the Levenshtein distance. In order to choose the more significant tag in a group, preference is given to words that can be found in the WordNet base. The last step of the cleaning procedure is to eliminate tags with a low frequency, or appearing only isolated.

In the context of library metadata, [Van Hooland et al. \(2013\)](#) describes as a first step for metadata reconciliation “profiling and cleansing of metadata”. Using an open source tool, authors describes cleaning activities such as deduplication (remove duplicate entries), atomization (explode overloaded fields), applying facets and clustering.

4.1.4 Metadata Reconciliation

On the metadata context, reconciliation refers to the process of finding a correspondence for some text string in a controlled vocabulary, thesaurus or ontology. To the extent of our problem, we are going to analyse strategies for mapping possible multi-language tags into defined ontologies, in order to be able to semantically process these tags.

The reconciliation approach described by [Van Hooland et al. \(2013\)](#) consists simply in searching the categories in pre-defined ontologies such as LCSH and Powerhouse Museum. This approach is followed by some content specific processing in order to equalize plurals.

[Lawler et al. \(2012\)](#) developed the Open Reconcile tool, a reconciliation tool tailored to help curators to ensure the compliance of datasets with controlled vocabularies. Alongside the automatic procedures, user are allowed to build a synonym table in order to give manual input to the algorithm.

A whole Semantic Tagging system is proposed by [Marchetti e Rosella \(2007\)](#). The system, implemented as a browser plugin, allows users to tag web resources and choose a corresponding semantic resources from knowledge bases such as Wikipedia.

As we can see, several conventional approaches do not include any semantic intelligence on the reconciliation task. This is not the case of the technique described in [Angeletou \(2008\)](#). In this case, author first performs a sense disambiguation, which consists of calculating the similarity distance to co-occurring tags, and then select the sense with the smaller distance. This procedure is deeper detailed in [Angeletou, Sabou e Motta \(2008\)](#). The second step is called Semantic Expansion, which is justified by the sparseness of the Semantic Web. In this step, synonyms and synonyms of the hypernym of the correct sense are included in order to search for semantic web entities (SWE). The process is finalized by searching for SWEs in the Watson platform, and choosing the most adequate according to the defined criteria.

Instead of grouping tags using semantic criteria, [Specia et al. \(2007\)](#) uses a statistical approach for this. An $N \times N$ co-occurrence matrix is built, where N is the number of distinct tags, and each element m_{ij} represents the number of times that tags i and j co-occur in different resources. Thus, the lines or columns of this matrix are vectors representing the tags, and the angular distance between them are calculated in order to cluster the closer tags. After building the clusters, terms are pairwise searched in ontologies in order to find the appropriate semantic entity. This procedure is also used for finding relations between the tags, which will be discussed in the following section.

In this section, several approaches make use of similarity measures. This topic is very extensive, and several strategies can be found on the literature ([HARISPE et al., 2015](#); [HARISPE et al., 2014a](#); [TRILLO et al., 2007](#); [CILIBRASI; VITANYI, 2007](#)). It is worth highlighting a paper by [Cattuto et al. \(2008\)](#), where several measures of relatedness are compared in the context of tags in social bookmarking systems. Relatedness is considered to be a special case of similarity, which is grounded only in the folksonomy (and not in external sources, as showed in [Angeletou \(2008\)](#)). The alleged reason for grounding the measures only in the folksonomy is the use of community specific terms, which may not be present in external vocabularies. [Cattuto et al. \(2008\)](#) presents three groups of relatedness measures: co-occurrence, distributional measures and FolkRank, which uses a similar approach as the PageRank algorithm.

4.1.5 Structure Emergence

Finding semantics entities related to tags is an important step. However, in order to build a knowledge base, it is necessary to find and qualify relations between these entities. Some of the above cited works also proposed strategies for this step.

[Specia et al. \(2007\)](#) searches if a pairs of tags appears on the same ontology, and in case of success, relations are extracted directly from the ontology.

A number of works, such as ([LIMPENS; GANDON; BUFFA, 2013](#)), use the WordNet database in order to determine the relation between two words. The hierarchical structure of WordNet allows to determine broader and narrower relations, as well as to calculate the distance between word through the WordNet tree.

A very interesting point-of-view on this topic is brought by [Limpens, Gandon e Buffa \(2013\)](#). In this work, a complete model for the semantic enrichment of folksonomies is presented including a socio-technical approach for managing diverging points of view, e.g., “Kevin agrees with the fact that soil pollution is a more specific term than pollution but Alex disagrees”. Figure 9 shows the proposed model. After driving an automatic reconciliation and structuring strategy, which is then validated or corrected by users, the divergences are managed by a conflict solving module.

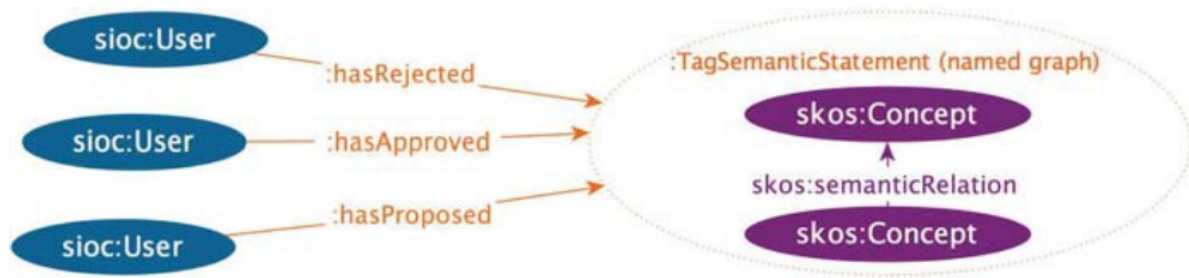


Figure 9 – SRTag RDF schema (LIMPENS; GANDON; BUFFA, 2013)

4.1.6 Automatic semantic tagging

Although this is not the main objective of this work, it is worth mentioning some strategies for automatic semantic tagging of documents. Allahyari (2016) proposes a probabilistic model based on DBPedi hierarchical model to automatically determine categories to documents. The model was successfully tested on a Wikipedia sample and on a Reuters database. Since categories are DBPedia resources, they can be considered as semantic metadata for linking purposes.

Chemudugunta et al. (2008) proposes a similar approach, but using unsupervised statistical learning. The generic model can be used both with human-defined concept and data-driven topics, and was tested against an educational text corpus.

4.1.7 Semantic Lifting in ODPs

The problem of semantic lifting in ODPs was tackled by Ermilov, Auer e Stadler (2013b) and Ding et al. (2011a). In Waal et al. (2014), a strategy for lifting datasets in ODPs to the Linked Data cloud is presented. In all these works, however, the semantic lifting refers to the datasets, and not to metadata.

4.2 An analysis of metadata in ODPs

Besides having an overview about literature related to semantic metadata, it is also necessary to the proper development our work to profile the use of metadata in Open Data Portals. In order to propose innovations, it is mandatory to know the main problems of real-world metadata usage.

In this section, we profile the use of metadata in Open Data Portals, with a special focus on tags. The analysis is restricted to systems running CKAN², the standard open-

² Available at <<http://ckan.org>>

Table 8 – Summary of data used in the experiment.

Portals	140
Analysed Portals	87
Tags	290,075
Groups	1,701
Datasets	470,551
Datasets without group	417,393
Datasets without tag	172,157

source software for ODPs. The CKAN community publishes a census³, where 139 portals were listed at the time of the experiment. Through the API offered by CKAN, we tried to obtain data from all portals, but only 87 responded adequately when the assessment was performed (March of 2016). Reasons for the lack of availability were mainly that the portal was completely offline, the API was disabled or not responding at the same URL of the website or the portal was using an outdated version of CKAN.

The majority of ODPs is related to governments and public administrations at local, regional, national or continental levels. Some of them are also focusing on specific themes, such as energy or geothermal data. Although most portals are authoritative and run by governments and public administrations, some of them were built as civil society initiatives. A complete list of the analysed ODPs is available online⁴.

The analysed ODPs are quite heterogeneous. The number of datasets in each portals varies from 4 to 194,592, and the number of tags, from 8 to 59,208. Regarding the quality of the portals, although there is no general benchmark, *Open Data Monitor* attests a high heterogeneity within European ODPs. An informal quality assessment using the Five Stars of ODPs (COLPAERT et al., 2013) also shows that portals vary from simple data registries (one star) to a common data hub (five stars).

A summary of the experiment data is shown in Table 8. The code used to collect and analyse the data is available as an open-source project⁵.

The analysis is divided in two groups: local metrics, to analyse the quality of tags in a particular ODP, and global metrics, looking at the interrelations between portals, and with the Linked Open Data (LOD) cloud.

Regarding the other main tool for organizing ODPs – groups – Table 8 also shows the number of groups per portal, and the number of datasets inside each one. While the

³ Available at <<http://ckan.org/instances>>

⁴ <<http://bit.ly/1NGygtk>>

⁵ <<https://github.com/alantysel/StodAp>>

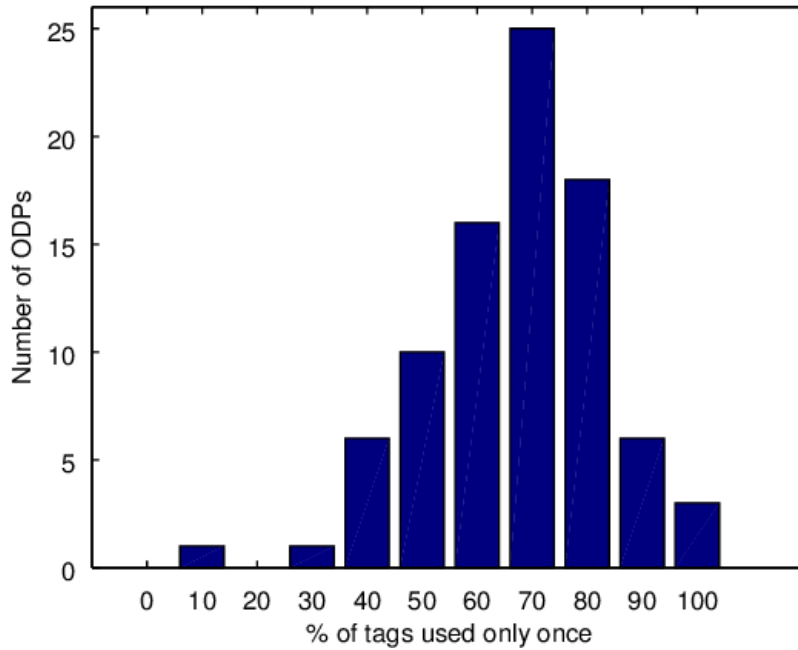


Figure 10 – Re-use of tags inside a portal. The graphic shows the distribution of the percentage of tags used only once.

tags are attributed to an average 3.88 datasets, groups contain a mean value of 67.45 datasets. This makes groups less selective than tags, which justifies our decision to focus on tags in this work. Moreover, while all 87 portals use tags, 18 do not use groups to organize data.

4.2.1 Local Metrics

4.2.1.1 Tag Reuse

The objective of this metric is to assess whether a single tag is being used to characterize several datasets, just a few or even only one. Creating new tags for each dataset can be considered a bad tagging practice. If tags are reused for several datasets, tag-based information retrieval will be more effective. Figure 10 shows the distribution of the percentage of tags used only once for each portal. The graphic shows a peak around 70% of the tags used only once. From the 87 portals, 75 use more than 50% of the tags only once. As a conclusion, tag reuse can be considered very low, thus effectively preventing the tags to be a suitable means to improve navigation, exploration and retrieval of datasets from ODPs.

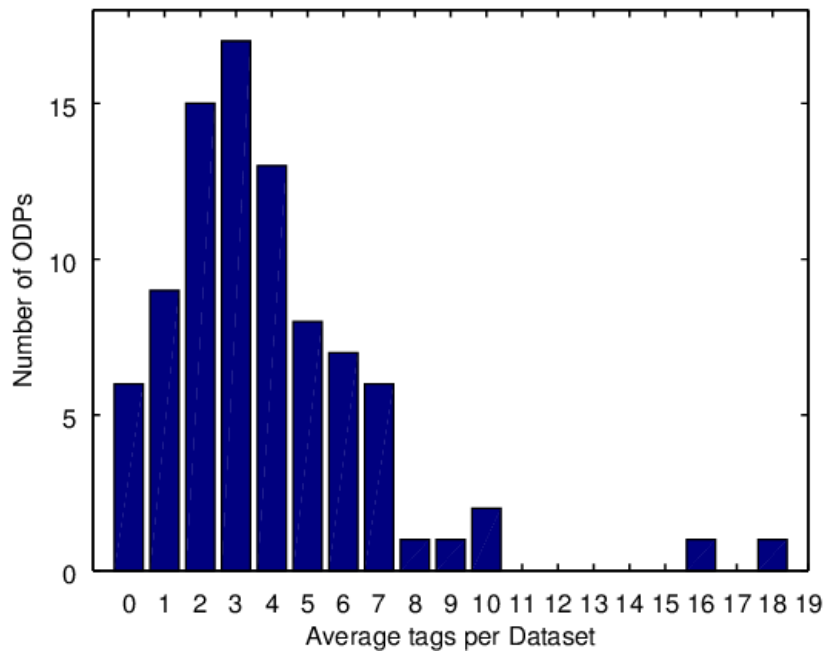


Figure 11 – Distribution of the average number of tags used per dataset in Open Data Portals.

4.2.1.2 Tags per dataset

This metric assesses the number of tags used per dataset. The goal is to verify, as in (UMBRICH; NEUMAIER; POLLERES, 2015), if the tag metadata is being actively used in the portals. However, the results of this metric cannot lead to further conclusions, since there is no optimal value for the number of tags per dataset. Using few and consistently used tags may support the organization of datasets better than many incoherently used ones. On the other hand, few tags may not label the content adequately. Figure 11 shows the distribution of the average tags per dataset for each portal. We can see that most ODPs apply between 1 and 7 tags to each dataset, with a peak around the value of 3. In general, we can affirm that describing datasets with tags is a common procedure in ODPs.

4.2.1.3 Tag similarity

By looking at the ODP tags, one can readily recognize that many tags differ only on capitalization, accents or singular and plural forms. Thus, this metric assesses whether several tags are being used with the same meaning. While recognizing these cases is easy for humans who understand the language of the tags, an automatic discovery of tags with the same meaning is not always straightforward. A simple approach is to convert the tags to lowercase and unaccented strings for comparison. Despite its simplicity, this method catches a significant number of cases such as `birth` and `Birth`.

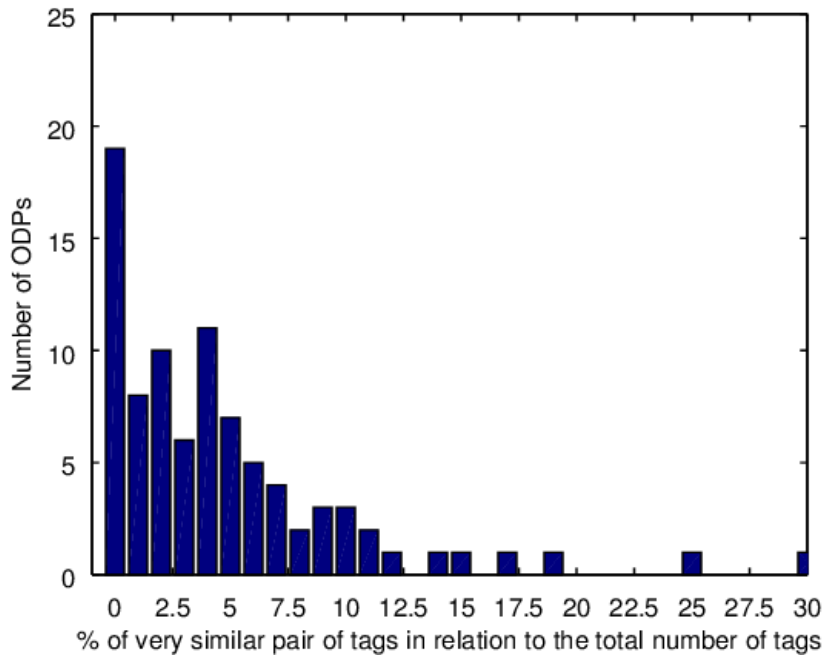


Figure 12 – Proportion of similar tags in ODPs, where the difference lies only capitalization or special characters.

A second possibility is to use the well known Levenshtein edit distance, which can also be suitable for detecting gender and plural differences, in some languages. This algorithm calculates the minimum number of character modifications – insert, delete and edit – necessary for turning a sequence into another. However, this method fails with tags containing numbers. For example, the Levenshtein edit distance between `budget-2010` and `budget-2011` is the same as between `Access` and `access`. Semantic-oriented methods, as detailed in [Harispe et al. \(2015\)](#), could also be used to detect synonymous tags.

Figure 12 shows a distribution of the percentage of similar tags inside each ODP. Similarity was checked using the simple approach. The occurrence of a significant rate of similarity reveals that there are few portals adopting a systematic tagging procedure. Despite the low percentage for some portals, in many of them similar tags still occur. Only 20 portals, out of overall 90, revealed no similar tags at all. It should be noticed that these portals use far less tags (148 per portal) than the average of all portals (2451 per portal), which may also be a sign of careful tagging.

4.2.2 Global Metrics

4.2.2.1 Coincident tags between portals

Different ODPs, especially governmental ones, can publish related data, which may also be tagged similarly. A similar measurement was used in ([UMBRICH; NEUMAIER;](#)

[POLLERES, 2015](#)). Using the same tag comparison approach as described in the local tag similarity metric, we found that 79,882 tags appeared in more than one ODP, which represents 28% of the total tags. This figure, however, should be carefully analysed. If we are interested in datasets from different ODPs tagged similarly, an overestimation bias may come from the fact that some portals act only as datasets harvesters, replicating the same datasets (and related tags). On the other hand, because portals are available in several languages, different tags could have the same meaning in different languages, what in turn tends to be an underestimation bias. In any case, the figure clearly indicates that there exists great potential for linking tags between open data portals. In fact, with this metric, our aim is to justify and motivate the development of a semantic tag curation approach for open data portals, which will be described in [5.3](#).

4.2.2.2 Tag expressiveness

A way of taking the tagging process one step further is to associate tags with resources or terms openly described in knowledge bases. In ([PASSANT, 2008](#)), while building the MOAT ontology⁶, authors designed the association of tags with meanings, represented by one or more URIs in the LOD cloud. With this expressiveness metric, our aim is to check if a tag is suitable to be connected to the LOD cloud, i.e., if there are candidate resources to represent its meaning.

Several knowledge bases are available on the Web, with DBpedia and WordNet being the most prominent ones. They are characterized by providing both a model of data organization – ontology – and the individual instances. DBpedia⁷ is build after Wikipedia knowledge base, and contains more than 38 million things, described in 125 languages using DBPedia Ontology.

WordNet ([FELLBAUM, 1998](#)) is one of the most used lexical database for the English language. Its strength relies on synsets describing the semantical relations between several senses of words.

In our tests for matching tags with semantic resources, we found that [<Lexvo.org>](#) ([MELO, 2013](#)), was the better service to search connections to different semantic knowledge bases, in several languages. Lexvo.org is connected not only to Wikipedia and WordNet, but also to Gemet, Wikitionary, Eurovoc, Agrovoc, OpenCyc and others. By providing an isolated term (in our case, the tag) and its language, Lexvoc.org returns the corresponding translations, as `lexvo:translation`, and if the term is English, it returns semantic resources, either as `rdfs:seeAlso` or `lexvo:means`.

Table 9 shows the results. The majority of tags (68.38%) did not correspond to any semantic resource according to this method. 8.15% of the tags were not evaluated either

⁶ <http://muto.socialtagging.org/mirror/moat.rdf>

⁷ <http://wiki.dbpedia.org/>

Table 9 – Expressiveness of tags. Percentage of main tags that could be associated to semantic resources. The tag universe considered here refers to clean tags, as described in Subsection 5.4.1, and represents 60.58% of overall tags.

	Absolute Occurrence	Weighted by Usage
Associated to a meaning	26.35%	36.06%
Not associated to a meaning	73.65%	63.94%

because they contain numbers, or because their length was equal or smaller than three. In those cases, results are mostly wrong. For 23.46% of the tags, at least one meaning or equivalent term was found, and their use represent a similar magnitude of 23.71%. Some tags can return several meanings, such as *leaves*⁸, for example: abandoning something, handing something to someone, or the plural of leaf, among others. In those cases, a further disambiguation procedure is needed.

It is not possible to guarantee that all associations were meaningful, and even worse, that the meaning intended by the tagger was correctly captured. The tag language was estimated by the ODP locale metadata, which can also be a source of errors if not correctly set. Some portals are also multi-language, and this characteristic is normally described. Further evaluations are needed in order to estimate the potential that ODP tags have to be connected to the LOD cloud. However, we see that at least one fifth of the tags correspond directly to a semantic resource. Providing context and a stemming pre-processing would probably enhance this result. Thus, we can say that some semantic potential is present on the tags.

After this analysis, we can affirm that: (i) tags in ODPs are widely used, but in a non-systematic way, which hinders their capacity of supporting information retrieval, and (ii) there is a potential for using these tags as connecting elements between ODPs, and for raising semantics from them. Next, we describe our proposal based on these statements.

4.3 Our contribution regarding the state of the art

In this chapter, we presented the state of the art in each of the steps for dealing with the problem of semantic organization of ODPs. In relation to the cited works, we are advancing on this field in the following aspects:

- Instead of dealing with folksonomies, i.e., several users tagging the same resources in the same platform, our problem is related to few administrators tagging different

⁸ <<http://www.lexvo.org/page/term/eng/leaves>>

resources on different platforms.

- Our approach is multi-language;
- We implemented a platform for semantically linking datasets through global tags;
- Dealing with the context of ODPs.

Offline: [Marchetti e Rosella \(2007\)](#) ([PASSANT, 2008](#))

5 Semantic Tags for Open Data Portals

As observed in the previous chapter, literature related to semantic enhancement of ODPs metadata has still some significant gaps as:

- Emerging semantics from the ODP context;
- Dealing with multiple languages;
- Tags attributed by few users, in a non-folksonomy style;
- Integrating multiple domains.

In order to tackle this issue, we describe in this chapter the Semantic Tags for Open Data Portals - STODaP approach for improving tag curation within and across ODPs, and for linking ODPs through its metadata.

Our main contributions are:

- A comprehensive analysis of tag usage in 87 ODPs, which justifies the need and benefits of better tools for managing tags;
- An approach for cleaning and reconciliation of tags in ODPs; and
- An approach for collaboratively connecting ODPs through meaningful shared tags.

In the first section, some considerations about metadata in open data portals are derived. In the following, the different concepts of tagging are put into perspective, in order to characterize tags in ODPs. Section 4.2 presents an analysis of the use of tags in several Open Data Portals, both from government and civil society side, and from various countries and languages. The main part of this chapter lies in Section 5.3, where our approach for semantic tags in open data portals is explained. Following sections presents some aspects about the implementation, the validation of the approach, and a conclusion.

5.1 Motivation

Analysing large amounts of data plays an increasingly important role in today's society. However, new discoveries and insights can only be attained by integrating information from dispersed sources. Despite recent advances in structured data publishing on the Web (such as RDFa and the schema.org initiative) the question arises how larger datasets can be published and described in order to make them easily discoverable and facilitate the integration as well as analysis.

One approach for addressing the problem of data dispersion are data catalogues, which enable organizations to upload and describe datasets using comprehensive metadata schemes. Similar to digital libraries, networks of such catalogues can support the

description, archiving and discovery of datasets on the Web. Recently, we have seen a rapid growth of data catalogues being made available to the public. The data catalogue registry datacatalogs.org, for example, already lists 285 data catalogues worldwide.

Data catalogues where data is supposed to be open, at least in the licensing sense, are usually called Open Data Portals (ODPs). Implementations that show the increasing popularity of ODPs can be seen, for example, in open government data portals, data portals of international organizations and NGOs, as well as scientific data portals.

In order to increase transparency and citizen engagement, governments and public administrations all over the world are implementing ODPs. These ODPs comprise large amounts of structured data, mostly in the form of tabular data such as CSV files or Excel sheets. They aim to be a one-stop-shop for citizens and companies interested in using public data produced by governments or civil society organisations. Examples are the [US' data portal](#), the [UK's data portal](#), the [European Commission's](#) portal as well as numerous other local, regional and national data portal initiatives.

In the research domain ODPs also play an important role. Almost every researcher works with data. However, quite often only the results of analysing the data are published and archived. The original data, that is ground truth, is often not publicly available thus hindering repeatability, reuse as well as repurposing and consequently preventing science to be as efficient, transparent and effective as it could be. An example of a popular scientific open data portals is the [Global Biodiversity Information Facility Data Portal](#). Also many international and non-governmental organizations operate ODPs such as the [World Bank Data Portal](#) or the data portal of the [World Health Organization](#). Although being a relatively new type of information system first commercial (e.g. Socrata) and open-source (e.g. CKAN) data portal implementations are already available.

Despite its recent popularity, Open Data and Open Data Portals still face significant impediments, as richly described in Section 2.8. Authors collected 118 socio-technical impediments for use of open data from interviews, workshops and literature. Some cited impediments were “absence of commonly agreed metadata”, “insufficiency of metadata”, “the lack of interoperability” and “difficulty in searching and browsing data”, showing that a great challenge for ODPs is the organization of data.

The open data organization challenge can be subdivided into two aspects: 1) structuring and organizing the datasets themselves and 2) providing well-structured and organized metadata for the datasets. The first aspect was, for example, tackled by approaches for semantic lifting of data by ([ERMILOV; AUER; STADLER, 2013a](#)) and ([DING et al., 2011b](#)), who tried to build general strategies for putting large open government datasets in the Link Data cloud. For the standardized structuring metadata,

the Data Catalog Vocabulary (DCAT)¹ (CYGANIAK; MAALI; PERISTERAS, 2010) was developed. However, the cross-portal metadata alignment and reconciliation can not be addressed by DCAT.

The metadata used to organize datasets in an ODP comprises categories or groups and most importantly labelling with free-text words or sets of words – the tags. The concept of tagging became popular within Web 2.0 services and aggregation tools like del.icio.us. The main advantages of tagging are the ease of classifying, and the crowd effect – resulting in the so called folksonomies – because all users were allowed to tag and share their contents. Tagging datasets in an ODP cannot be considered as folksonomies, because the process is mainly driven by portal managers and data publishers, and not by the actual users. As a result of this, the structuring effect of crowd-tagging and folksonomies is missing in ODPs.

A quick look over some ODPs reveals that most of them suffer from a very confusing organization of datasets. The first level of categorization uses the concept of groups. In general, they are stable and meaningful, but normally contain a large number of datasets. A more detailed classification should be done via tags, whose use in ODPs has the following issues:

- *Synonyms*: In most ODPs, there exists large number of synonymous tags, e.g., **crops** and **seeds**;
- *Different spellings of the same word*: Several tags are incorrectly written, or have differences in capitalization or accents, e.g., **baden-wuerttemberg** and **Baden-Württemberg**;
- *Lack of relationships*: There is no explicit relationships between the tags, e.g., **Community Centres** is clearly a specialization of **Community**, but this is not explicit;
- *Ambiguity*: As tags are written as pure text, ambiguity is prevalent in ODPs, e.g., the tag **apple**, which could refer to the fruit or to the company; and
- *Incoherence*: Tags do not allow any connection between different portals that use the same or equivalent tags, e.g., two datasets tagged with **budget** in different portals are not connected.

As a result, the navigation, exploration and search within individual, but in particular also across ODPs is significantly hampered.

5.2 Definition of an Open Data Portal

According to Colpaert et al. (2013), an Open Data Portal is “a collection of systems set up to make Open Data used and useful”. A formal definition of an ODP can be found in Umbrich, Neumaier e Polleres (2015). However, in that case, the focus is general

¹ Available at <http://www.w3.org/TR/vocab-dcat/>

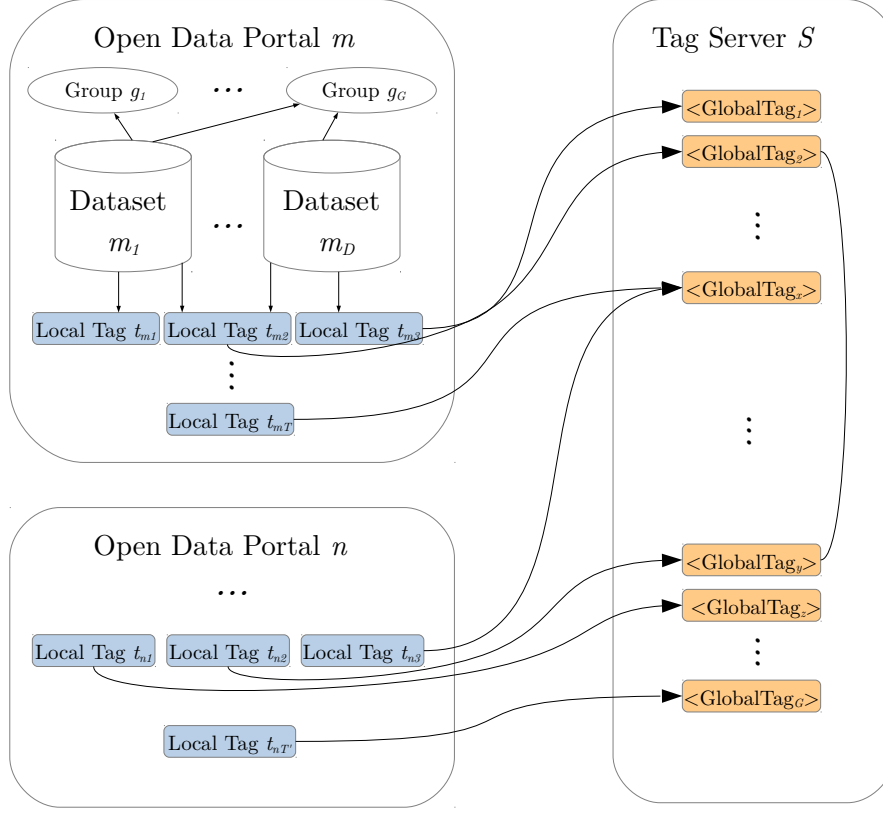


Figure 13 – Relevant elements of the Semantic Tags for Open Data Portals system.

metadata analysis, which turns their definition unsuitable to be used here. In this section, we will describe the elements that are present in the context of this work.

Figure 13 shows the relevant entities and relations that are used in the remainder of this paper. An *Open Data Portal*, in this context, is a collection of datasets, which hold open data resources online. *Datasets* can be organized in *Groups*. Each *Dataset* belonging to an ODP can be tagged with local tags. Each *Local Tag* also belongs to an ODP, and can be used to tag one or more datasets. In this architecture, local tags are connected to *Global Tags*, stored in a collaborative *Tag Server*. Several local tags from different ODPs can be associated to a single global tag, which can also have semantic relationships with other global tags.

5.3 Overview of the STODaP Approach

In this section, we give an overview on the tag reconciliation approach for cleaning up and connecting ODPs, supported by software tools both at the local and global contexts.

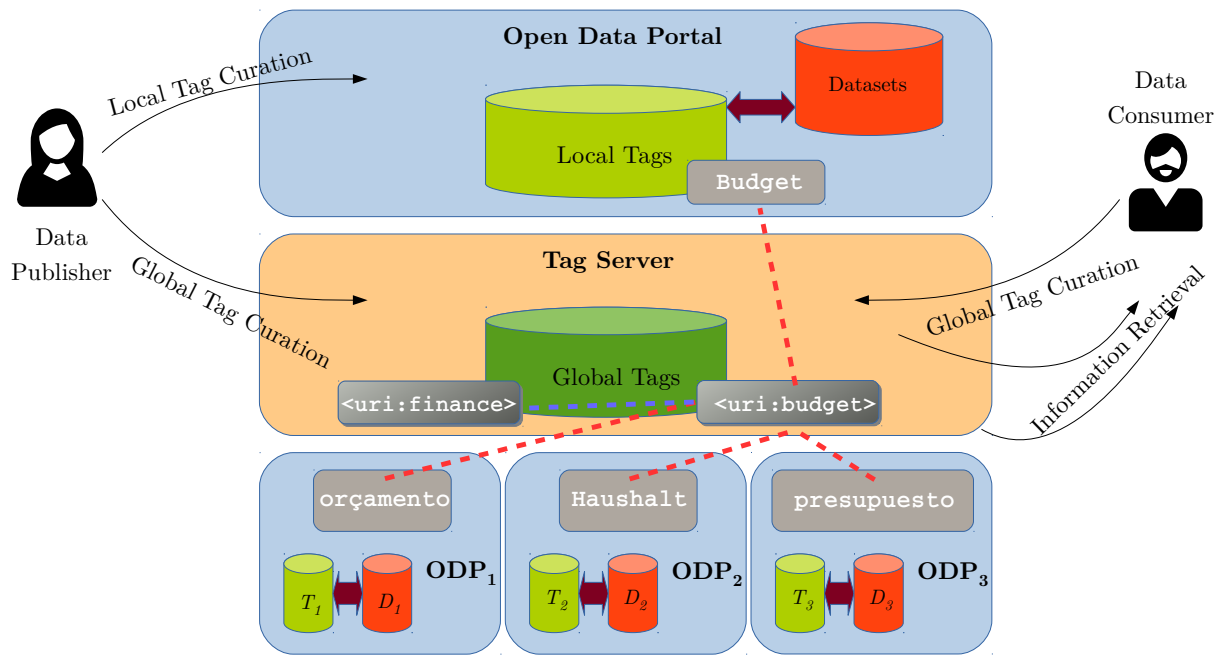


Figure 14 – Overview of the StodAp approach. Local tags are connected to a corresponding global tag within a central tag server. Data managers responsible for ODPs may use tools for local tag curation, as well for maintaining the tag server. This task is also expected to be performed by data consumers.

The objective of the approach is to tackle the main problems identified by the metrics described in the previous section, and thus to facilitate data organization and linking through metadata descriptions of ODPs.

Figure 14² shows an overview of the proposed approach. Data publishers in charge of ODPs are offered tools for local tag curation. These tags are then connected to global tags hosted in a central Tag Server, which can be collaboratively edited both by data consumers and publishers. They can add new semantic descriptions to the global tags, establish relations between them, and also create new links between global and local tags. Data consumers have the option to retrieve data directly from ODPs, or through references gathered from the central server.

The following three sections are dedicated to further detail the approach. Section 5.4 explains the procedure used to build the first version of the semantic tag server. In the following, Section 5.5 describes the STODaP tools and methodologies for open data managers and users in order to maintain the tag server. Finally, Section 5.6 describes the implementation of the tools.

² Icons by SimpleIcon from www.flaticon.com are licensed under CC BY 3.0.

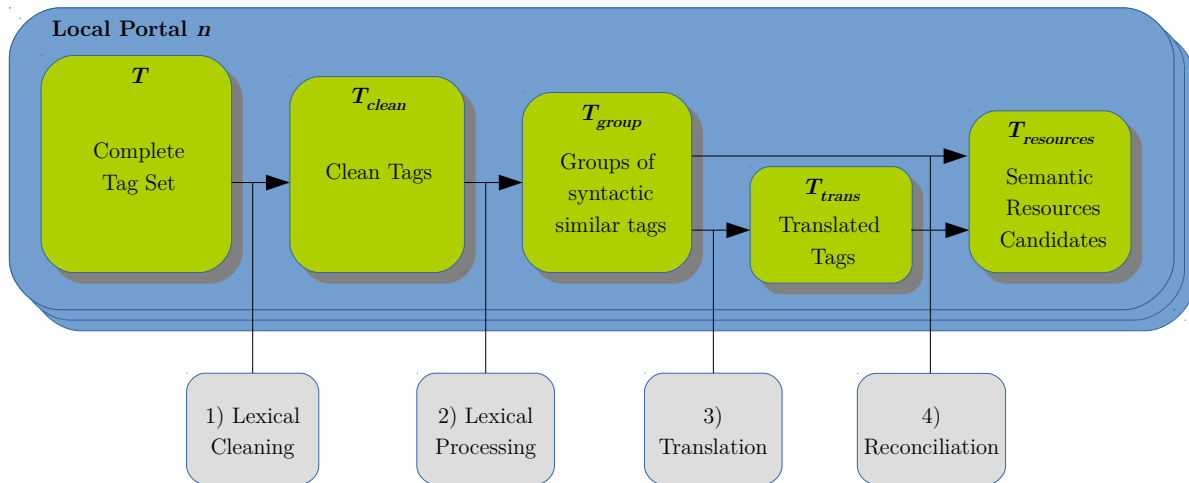


Figure 15 – Overview of the local tag processing.

5.4 Building the STODaP server

In order to build the first version of the STODaP server, a tag harvesting was done through 87 ODPs. Almost 300.000 tags were processed, using their names and informations as language and groups that the tagged datasets belong to. In the following, we describe first the procedures for the individual portals, and then the procedures for whole tag set until we come to the structured Global Tag Set.

5.4.1 Local Processing - Clean Up and Reconcile

An overview of the procedure applied locally, i.e., for each ODP, is shown in Figure 15. In the figure, each green block represents a processing phase, which is materialised in a set of tag representations. The grey blocks describe the transformations suffered by the tag sets from one phase to another. The aim of the local processing steps is to transform freely written tags into semantic resources that are candidates for representing the datasets they are associated. Each transformation applied to the tags on the local processing is detailed below:

Lexical Cleaning: The complete tag set T is the set containing all original tags found in one portal. Firstly, the *Lexical Cleaning* is applied in order to discard tags with low probability of getting a semantic meaning. At this point, some heuristics are applied, and a tag is discarded if it is:

- smaller than 4 characters;
- composed of numbers and alphabetic characters;
- exclusively composed by uppercase characters;
- not started by a alphanumerical character;

- larger than 5 words; or
- not applied to any dataset.

Lexical Processing: After the Lexical Cleaning, we have the resulting set T_{clean} of clean tags, with a higher probability of being reconciled with semantic concepts in ontologies. The following procedure is the *Lexical Processing*, which aims to group tags that have a lexical similarity. These similar tags have a high probability of representing the same meaning, with small lexical variations. In order to determine this similarity, we apply the Levenshtein edit-distance to the lowercased and unaccented tags (which means that `Açaí` will be transformed into `acai` before measuring the distance). Based on manual experimentation, we consider that tags with an edit-distance of 0 or 1 are similar. This distance captures plural, gender and temporal variations in most of the languages present in our sample. The proceeding results in the set T_{group} of syntactically similar tags.

Translation: The sample used to build this tag server contains portals in 22 different languages. Thus, it is necessary to use translation services on the Web to transform words from their original language to the English language. English language was chosen because of the higher availability of translation services, and also because the main ontologies have their terms described necessarily in English, and possibly also in other languages. It also significant that 43% of the portals are in English (according to the provided metadata), and their tags represent 83% of all tags. Each group of similar tags from T_{group} was translated, resulting possibly in a set of translations for each group. The new set achieved in this processing is T_{trans} .

Reconciliation: The previous proceeding results in a set T_{trans} of groups formed by all the related translations. Until this moment, we were dealing with string of characters. In this stage, these names will be the input for searching semantic representations for the tags. In order to get the widest spectrum of possibilities, the search for semantic resources is done for all lexical representations of the tag, stored in T_{group} , and also all possible translations of it in T_{trans} . The resulting set will be denominated $T_{resources}$.

In order to illustrate the procedure, [Table 10](#) shows an example using real tags from the Brazilian Data.gov.br. From T to T_{clean} , tags containing numbers, too small or representing abbreviations were removed. Then, similar tags were grouped to form T_{group} . The translation process could not find an equivalent for the first group. Even so, the semantic search-engine was able to find a matching resource for `Acidente de trabalho` (accident at work), as well as for the other two.

It is important to notice that the process described above is subject to several failures. On the Lexical Cleaning step, meaningful tags with less than 4 characters may be discarded, as well as unintentionally uppcased words. On the Lexical Processing stage, it is possible that in some languages the same word starting with capital and non-capital letters have different meanings. With a higher probability, words differing

Table 10 – Examples of tags in each step of the procedure.

T	T_{clean}	T_{group}	T_{trans}	$T_{resources}$
Acidente de trabalho, Acidentes de trabalho, CNAE, finanças, Folha SA.23, Folha SB.23 município, orçamento, UF	Acidente de trabalho, Acidentes de trabalho, finanças, orçamento	{Acidente de trabalho, Acidentes de trabalho} finanças, orçamento	– finances, budget	{gemet:9366, eu-rovoc:825}, eionet:3194 eionet:1025

from edit-distance of 2 may also have different (or even opposed) meanings, such as **child-death** and **child-health**, found on data.gov.uk. On the Translation phase, the main problem lies on polysemy, where the same word has several meanings. While also heavily dependent on the translation tools, providing side tags or other metadata can help the algorithm finding the right translation. Finally, when searching for the meanings, there is a great dependency on the tool used and the available knowledge bases.

5.4.2 Global Processing - Interlinking Portals

After reaching the last stage of the local processing stage for each one of the 87 portals, a joint process starts over $T_{resources}$ in order to build the Semantic Tag Server. At the global processing stage, there are three main steps:

1. Select meaningful tags;
2. Create global tags and connect local tags to them; and
3. Discover and qualify relations between tags.

In order to accomplish this objective, we propose the global procedure shown in Figure 16.

Significance Selection: We start the global processing with a joint set $\mathcal{T}_{resources}$, which contains $T_{resources}$ from all portals. In this set, a *Significance Selection* process is driven, in order to determine tags that will be useful on the information retrieval process. This is an heuristics based process, which considers: (i) Success on finding semantic candidates for the tag; (ii) the number of datasets pointed by this tags; (iii) the quality of the semantic resources candidates.

Semantic Processing: After this step, the Global Tags will be derived. Global Tags are semantics entities, who have a main name in English, several translations, and point to local tags, which in turn connect to datasets located into Open Data Portals.

Table 11 – Examples of groups in some ODPs

Data.gov	Data.gov.de	Dados.gov.br	Data.buenosaires.gob.ar
Aging / Agriculture / Business / Climate / Consumer / Disasters / Ecosystems / Education / Energy / Finance / Health / Law / Local Government / Manufacturing / Ocean / Public Safety / Science & Research	Population / Education and science / Geography, Geology and the GEO-DATA / Laws and justice / Health / Infrastructure, building and housing / Culture, leisure, sport, tourism / Not yet categorized / Public administration, budget and taxes / Politics and elections / Social / Transport and traffic / Environment and the climate / Consumer protection / Economy and work	Municipal Chamber / trade, services and tourism / culture, leisure and sport / data sets in the spotlight / defence and security / economy and finance / education / public facilities / geography / government and politics / housing, sanitation and urbanism / health information / industry / justice and law / environment / person, family and society / management platform indicators / multi-year plan / international relations / health / work / transportation and transit	economic activity / public administration and policy / culture and recreation / education / infrastructure and public works / environment / mobility and transport / health and social services / security / urbanism and territory

Structure Emergence: Finally, relations between Global Tags in set G will be searched on the ontologies they appear in order give a structure to the Global Tag set G_{struct} . The first strategy is to search for relations on the reconciled ontologies, and set this relation between the global tags. Thus, relations as `skos:related`, `skos:narrower` and `skos:broader` can be set. However, at this point we notice need of an upper classification scheme.

The ODP model, as shown in [Figure 13](#), includes a Group element, to which one or more datasets can be associated. Thus, it is possible to consider that a tag associated to a dataset which is in a group is also related to this group. However, only 11% of all datasets in our sample are associated to groups, and only 13% of the tags are associated to datasets in groups.

If we look to some ODPs which are organized in groups, it is possible to see a similar organization. In [Table 11](#), we list the groups of 4 ODPs. If we look at the table, it is clear that in the context of open government data portals, there are some specific categories, but the portals also share common subjects, such as Health, Education or Culture.

Thus, after translating all the group names, we verified that 62 group names occurred in three or more portals. These were chosen as the first Global Groups. The

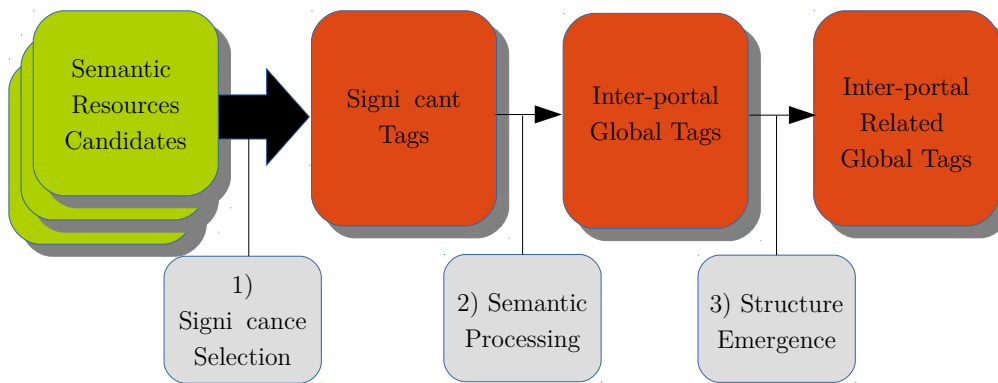


Figure 16 – Overview of the global tag processing.

second step consisted in verifying the lexical similarity between all groups and the Global Groups in order to associate groups with Global Groups. Some distortions were observed, such as **sport** being associated with **transport**, or **culture** with **agriculture**. These errors were manually corrected.

Finally, groups were reconciled with general-purpose ontologies. Particularly, the Gemet Thesaurus³ fits well for this purpose.

5.5 Use and Maintenance of the STODaP server

After building the Semantic Tag Server (STODaP), it is necessary to maintain and enhance the tag corpus alongside the evolution of ODPs, as well as to maintain the server updated. In this subsection, the strategy for it will be presented, locally and globally at the individual portal level, and at the server level.

5.5.1 Local Part - Cleaning up tags

Section 4.2.1 showed that ODPs suffer from low reuse of tags, and that there is a significant tags duplication due to slight spelling differences. In fact, both problems – low reuse and duplication – are connected, since merging similar tags improves tag reuse. However, low tag reuse can be also attributed to the lack absence of a standard tagging procedure, which would guide users in this task.

To address this problem locally at a particular ODP, we propose an approach for reconciliation of tags.

First, we offer three levels of semi-automatic tag merging strategies:

³ Available at <<http://www.eionet.europa.eu/gemet/>>

1. With high confidence, we suggest merging tags that differ only by capital letters or special characters. In many ODPs, this strategy will already achieve significant results, as shown in Figure 12.
2. After running the first strategy, the Levenshtein distance is computed for all remaining pairs of tags. Tags with distance one or two are suggested for merging, in order to catch plural/gender variations, such as **worker** and **workers**. However, false-positives like **widow** and **window** may appear. Tags composed only by numbers (to avoid merging tags representing years) or less than 4 characters are not included.
3. Finally, we use semantic measures (HARISPE et al., 2014b) to determine the semantic similarity between two tags. In this case, the tags **autumn** and **fall** have a high similarity, and thus will be suggested for merging.

It must be noted that all these approaches have originally quadratic time complexity, because every pair of tags has to be computed. However, sorting tags alphabetically turns the problem into linear in strategies 1 and 2 (however, with possible losses in 2), and ignoring tags without correspondence in dictionary reduces the dimension in strategy 3.

After this cleaning procedure, we offer users the opportunity to link each local tag to a global correspondent at the tag server, described in the sequel.

5.5.2 Global Part - Semantifying Tags

With the aim of building a common and collaborative basis for interlinking ODPs, we developed a Global Tag Server. The conceptual rationale is:

1. To assist individual ODPs enhancing the quality of their tags, by assigning a common agreed meaning to them;
2. To create a collaborative platform for meaningfully linking ODPs.

The Global Tag Server hosts the description of global tags. Each global tag may be associated to one or more Linked Open Data resources, representing their semantic meanings. Linking to the local tags is accomplished via the URIs which represent a local tag in its context. The global tags can also have several types of relations between each other, such as **skos:broader**, **skos:narrower** or **owl:sameAs**. Figure 17 illustrates the concept with an example.

The example shows the global tag identified by the URI `<http://stodap.org/tags/Budget>`. With this global tag, a meaning and some URIs of local tags are associated. The global tag is also semantically related to other global tags, using the SKOS vocabulary. The MUTO ontology⁴ is used to define some concepts and relations between the tags, like **muto:Tag**, **muto:taggedResource**, **muto:hasTag** and **muto:hasMeaning**.

⁴ `<http://muto.socialtagging.org/core/v1.html>`

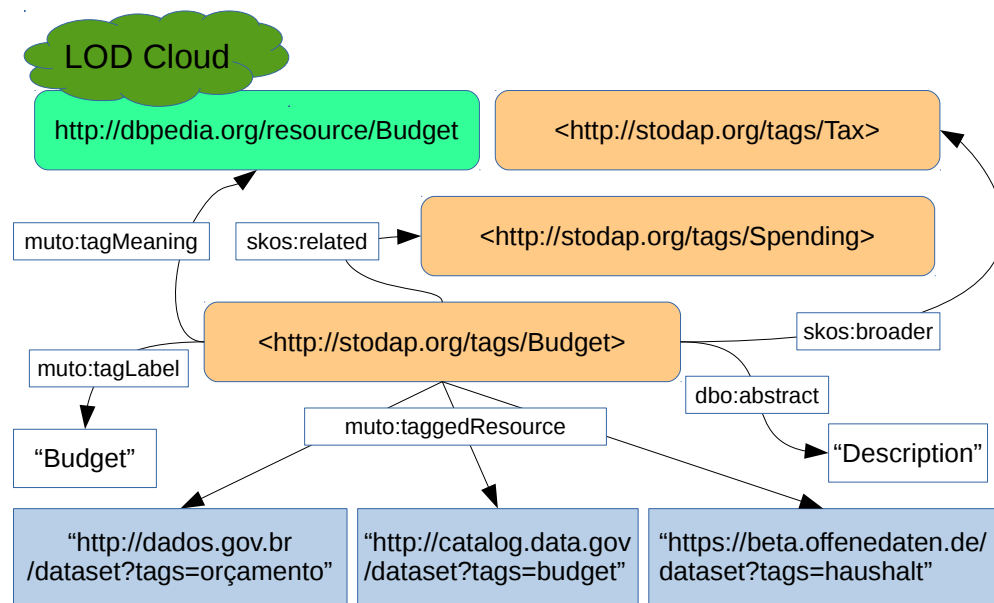


Figure 17 – Example of the STODaP model showing relationships of the global tag `<http://stodap.org/tags/Budget>`.

5.5.3 STODaP Server - Interlinking Portals

The first step for building the semantic tag server is to harvest metadata from open data portals. After the initial setup, a strategy for maintaining the portal up-to-date is needed.

Adding a new portal

When a new ODP is added to the system, a setup procedure is followed:

- Harvest tags, datasets and groups metadata;
- Clean and group similar tag;
- Translate tags, in case of non-English portal;
- Reconcile tags with existing Global Tags;
- If reconciliation is not successful, search lexvo.org and try to create a new global tag;
- Groups: reconcile with existing global groups or create new ones.

Updating a portal

When an ODP is updated, the procedure followed is:

- Harvest tags, datasets and groups metadata;
- Verify which datasets were inserted or modified

5.6 Implementation

In this section, we detail the technical procedures related to the previous sections. We first describe the implementation of the procedure for discovering the Global Tags, discussed in [Section 5.4](#). Then, we describe the implementation of two CKAN plugins: (i) *CKAN Tag Manager*⁵ and (ii) *CKAN Semantic Tags*⁶, which materialize the ideas reported in [Section 5.5](#). Finally, we describe the implementation of the Semantic Tag Server for Open Data Portals.

5.6.1 Building Global Tags

- Harvest all tags from portals
- Filter significant tags
- Group syntactically similar
- Translate - lexvo + yandex
- Reconcile - lexvo - several ontologies
- Choose gemet tags and create global tags and links
- Create global groups, and reconcile
- Associate global tags to global groups

5.6.2 CKAN Tag Manager Plugin

The CKAN Tag Manager one offers an environment for tag curation directly inside the CKAN platform. It comprises basic functions such as deletion and editing of tags, and advanced function aimed to enhance the quality of tags. In this sense, the plugin checks:

- Very similar tags, differing by capitals or special characters;
- Similar tags, with a Levenshtein distance ≤ 2 (after lowercasing and unaccenting)
- Possible synonyms, using Natural Language Toolkit ([BIRD; LOPER; KLEIN, 2009](#)).

In all those cases, the user is offered the option of merging the respective pair of tags. [Figure 18](#) shows a screenshot of the CKAN Tag Manager.

5.6.3 CKAN Semantic Tags Plugin

The CKAN Semantic Tags plugin implements the connection between a CKAN instance and the Global Tag Server. Each local tag can be associated to a global tag from the server. After the association, datasets linked with a local tag also point to the global server, as shown in [Figure 19](#).

⁵ <https://github.com/alantysel/ckanext-tagmanager>

⁶ <https://github.com/alantysel/ckanext-semantictags>

5.6.4 Semantic Tag Server

The tag server is implemented using the collaborative *MediaWiki*. Specially, the *Semantic MediaWiki* extension (KRÖTZSCH et al., 2007) is used in order to include properties and integrate the global tags in the LOD Cloud, through the export of RDF files. The page of a global tag is shown in Figure 20. Each global tag page is build using semantic templates and forms, in order to facilitate consistency and coherency and to be more user-friendly.

5.7 Results

We describe in this section some results achieved with the STODaP approach. At the global level, it was possible to implement the global tags server and to test the performance.

5.7.1 STODaP Server

In order to test the system, an open-source implementation of STODaP was created and deployed at <http://stodap.org>. The following approach was used create 663 global tags at the server:

- From the 220,567 tags harvested, we selected the 663 that were used in more portals, representing all tags used in 10 or more portals;
- Using the Lexvo.org service, we found URI candidates to represent the tag meaning via the `lexvo:means` property;
- Using the Lexvo.org service, we found translations and synonyms for the tags via the `rdf:seeAlso` and `lexvo:translate` properties;
- We searched for the translations and synonyms in the harvested tags and included the results as `muto:taggedResources`, together with the portals tagged with the original term;
- Using the [Natural Language Toolkit Library](#), we searched for semantic similar global tags, which were added as `skos:related`.

The occurrence of the original tags among the portals, and the results after including the translations and synonyms can be seen in Figure 21. The graphic shows the 30 most used tags, and the achieved increment in the number of relations. The occurrence of tags denoting years can also be noticed. Obviously these tags have no synonyms nor translations, and thus no increment is shown. It is also worth mentioning that the tag `test` is the fourth most used one. This fact is probably related to the early stage of development of some portals.

5.7.2 Local Level

At the local level, the main potential achievements are at the tag curation process. As shown in [Figure 12](#), a considerable number of pairs of tags differ only by capital or accented characters. Using the naive approach to merge similar tags in every portal would result in reducing the number of 14,168 local tags, which represents 6.4% of the total number of tags. Lowercase and unaccented tags differing by a Levenshtein-distance from 0 to 2 represent a total of 35,066 pairs, or 15.8% from the whole tag universe. However, as discussed above, this approach can lead to false-positives and thus requires manual checking.

5.8 Conclusions

In this chapter, we presented an approach for metadata reconciliation within and among Open Data Portals. The use of tags was analysed, and several problems were found, such as a low tag reuse rate and the overall existence of different tags for the same meaning. On the analysis we also found that several portals share the same tags, showing that tags have a good potential to be linking elements among datasets. Converting tags into semantic identifiers was also shown as a viable option, even though more sophisticated methods have to be investigated. Based on these findings, we derived the STODaP approach, which comprises two parts: a local one, aimed at cleaning up and enhancing the quality of ODPs tags, and a global one, for connecting ODPs through semantic tags. The implementation of both shows that significant enhancements can be achieved both at the individual ODPs and the global levels.

Future research and development includes a tag suggestion approach for ODPs which takes into account the related tags at the tag server, using collective knowledge as in [Sigurbjörnsson e Zwol \(2008\)](#). Using the possibly structured data of the ODPs in order to improve tagging suggestions is also a research direction that should be followed. At the global level, an interesting approach is to detect the emergence of schemas from the tags, as described in [Robu, Halpin e Shepherd \(2009\)](#). We will also call for the attention of the open data community in order further to advance collaborative strategies for enriching the tag server. For STODaP to realize its full potential, ODP administrators and users should be involved and (meta)data literacy needs to be improved.

The “openness” of open data is still limited by many factors, including politics, data literacy and technology ones. With this work, we contributed to the organization and interconnection of ODPs, and thus, given a step further on the enhancement of the open data field.

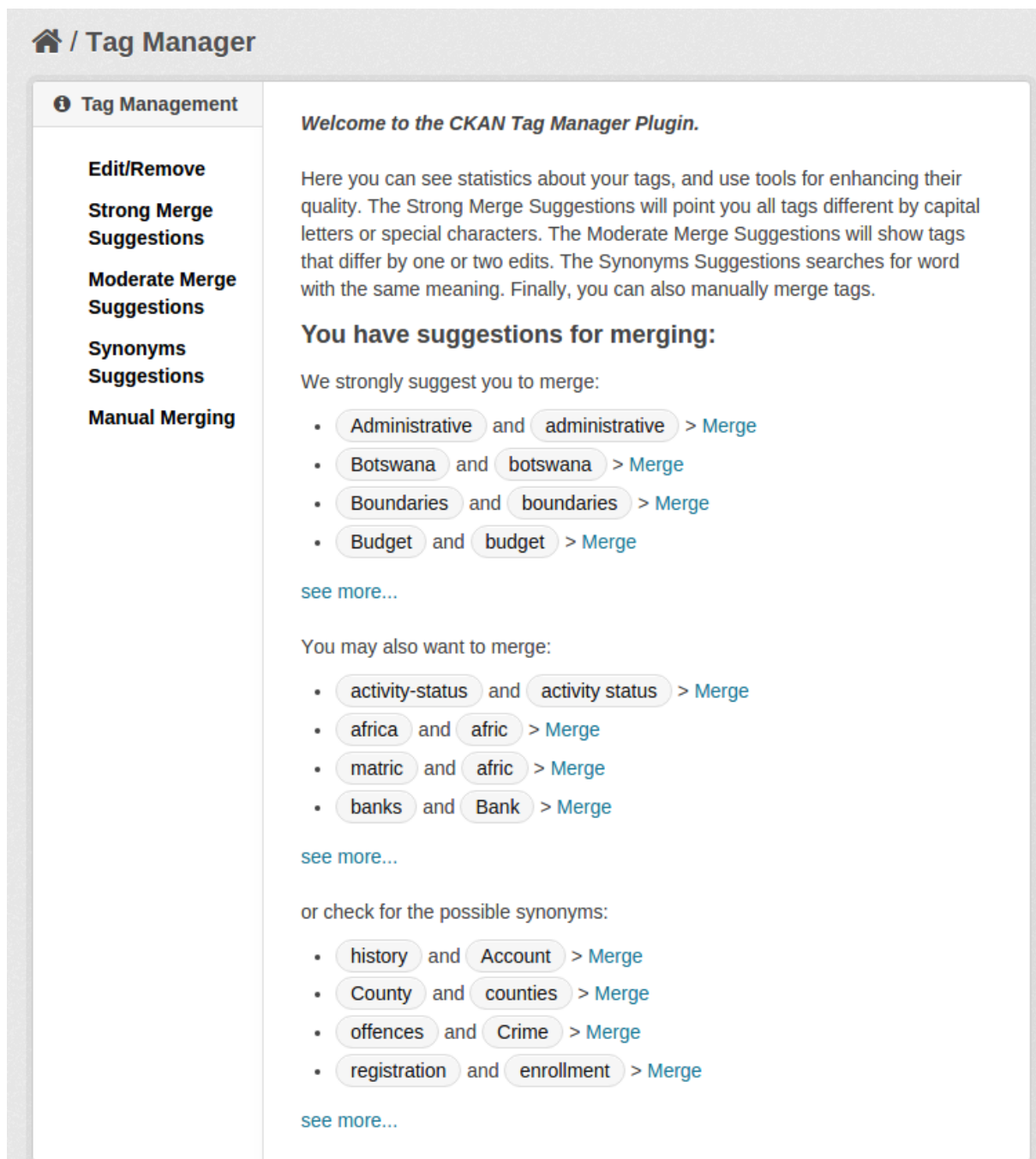


Figure 18 – Local tag curation in a CKAN instance. The plugin offers possibilities of manual and semi-automatic tag merging. The first block contains only valid suggestions, while the second block shows 2 false-positives. The synonym module also detected plurals. Tags in this example were extracted from the africaopendata.org portal.

Conjunto de dados
Grupos
Fluxo de Atividades
Relacionado

Preços de Alimentos

Índice de preços de alimentos, medido pela FAO - Organização das Nações Unidas para alimentação.


Dados e recursos


Food price index
The FAO Food Price Index is a measure of the monthly change in international...

Explorar

FAO
alimento > **Food**

Figure 19 – Detail of dataset in an ODP. The dataset is tagged with two tags, and one of them (*alimentos*) is connected to the global tag <http://stodap.org/tags/Food> through the `mutu:hasTag` property.

STODaP

Semantic Tags for Open Data Portals

Alan Talk Preferences Watchlist Contributions Log out

Page Discussion Read Edit View history More Search

Research

This is the page of the global tag Research. Below, you will find a description of the tag, some resources associated to it and the link for Open Data Portals tagged with a related local tag.

Abstract [\[edit\]](#)

Research comprises "creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications." It is used to establish or confirm facts, reaffirm the results of previous work, solve new or existing problems, support theorems, or develop new theories.

Open Data Portals [\[edit\]](#)

- opendata.admin.ch (Forschung)
- publicdata.eu (forskning)
- etsin.avointiede.fi (tutkimus)
- datahub.io (Research)

Resources [\[edit\]](#)

- http://www.fao.org/aims/aos/agrovoc/c_6513
- <http://en.wiktionary.org/wiki/research>

Related Tags [\[edit\]](#)

- [survey](#)
- [census](#)
- [projects](#)
- [service](#)
- [activity](#)

Category: **Muto:Tag**

Figure 20 – Semantic Tag Server for Open Data Portals. Simplified example of the global tag *Research*, which is linked to 48 Open Data Portals and 16 semantic resources. In the screenshot, we illustrate a few of them.

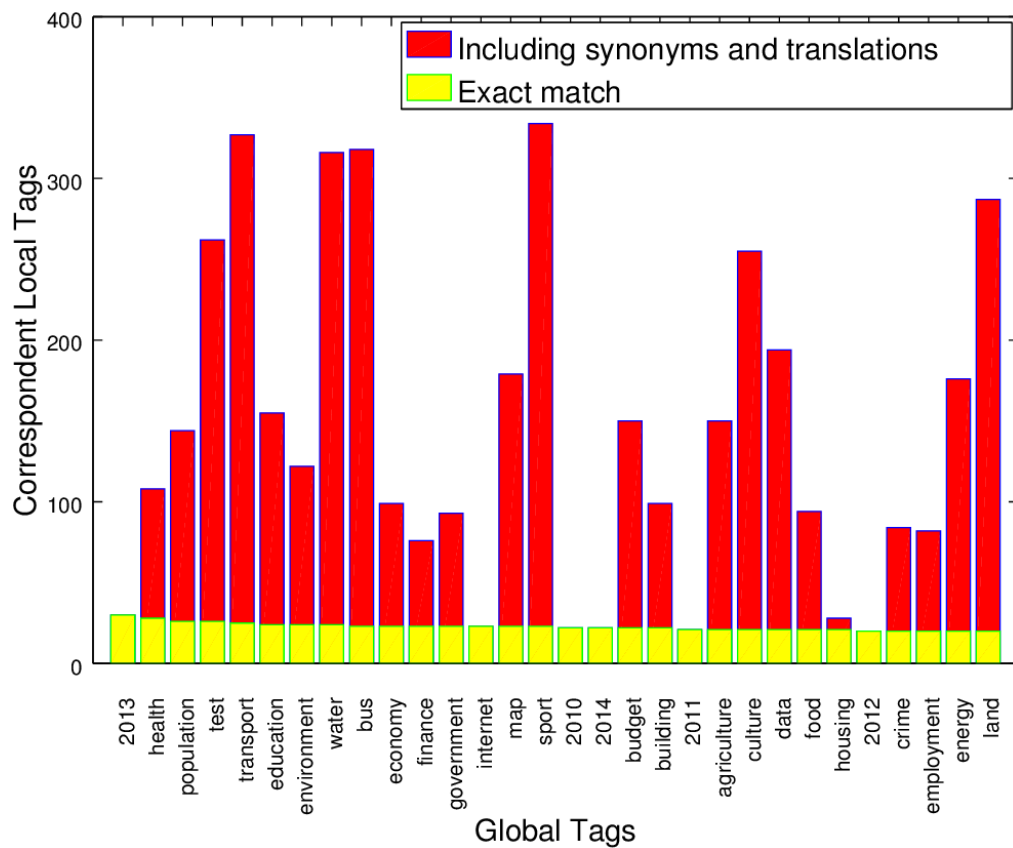


Figure 21 – Correspondence between local and global tags. The yellow bar shows the number of exact occurrences of the tag in ODPs. The red bar shows the improvement when considered translations and synonyms, which can also occur in a same portal. This explains the numbers over 90.

6 Evaluation

In the previous chapter, STODaP approach was presented in details, as well as the supporting tools and their implementation choices. Practical results were also characterized, showing concrete achievements on the open data organization problem.

In this chapter, the evaluation of STODaP server is described. The system was compared to other mechanisms on the task of searching for open datasets. We first present a theoretical background on search engine evaluation methodologies in [Section 6.1](#). Then, we show the experimental setup in [Section 6.2](#) and the results in [Section 5.7](#). Some concluding remarks are driven on the final section.

6.1 Methodology evaluation background

As the amount of online available data gets bigger and bigger, search methodologies are increasingly necessary to allow users accessing relevant content. Thus, it is crucial to develop evaluation techniques that allows researchers to compare different algorithms and find the most adequate ones for each context.

[Cheng, Hu e Heidorn \(2010\)](#) developed two measures for assessing *user satisfaction* and *user effectiveness* on Interactive Information Retrieval systems. The first one is called Normalized Task Completion Time (NT), and is calculated as the relation between task completion times for novices and experts. Following the same reasoning, the Normalized User Effectiveness (NUE) evaluates the relation between relevant documents retrieved by novices and experts, proportional to NT. Authors claim that this normalization procedure turns the measures more stable against task complexity variations. Results show that the NT is highly correlated to user satisfaction, while NUE is a better indicator for effectiveness when compared to simple task completion time. The learning curve was also better explained by NT and NUE than by task completion time.

In a contrary direction, [Xu e Mease \(2009\)](#) defend the use of task completion time as a robust measure to assess in which extent the search engine helps users to complete a task. Additionally, these authors found a negative correlation between user satisfaction and task completion time. An important result of this study is a mathematical development which shows that a cross-over design reduces significantly the variance of the experiment. Cross-over design means that, when comparing systems A and B on several tasks, every user tests both systems and completes all tasks once, half of them in A, and the other half in B.

In a survey dealing specifically with faceted search, [Wei et al. \(2013\)](#) presents a

review about relevance and cost-based metrics on the faceted search context. Regarding relevance metrics, authors go through a number of works which use precision, recall or F-measure in the same way as on non-faceted search evaluation. Cost-based metrics look at the time needed to complete a search task, and memory usage. These metrics were used to compare performance between faceted and non-faceted engines.

Although the Web and search engines have dramatically changed in the last 10 years, the perspective brought by Vaughan (2004) is still relevant. The focus in this work relies on the quality of ranking, i.e., the order in which results are presented. Both works presented previously rely on the task completion time, which brings with it factors that do not depend on the system, e.g., users ability, and factor not directly related to the search-engine, such as usability. By looking specifically at the ranking quality, the evaluation methodology may ignore these aspects, and keeps full attention on the search mechanism. In this work, author proposes non-binary counterparts to the traditional precision and recall measures, with the intention of adding human relevance judgement aspects to the evaluation. Specifically, two measures are proposed: (i) *Quality of result ranking*, as counterpart of precision and (ii) *Ability to retrieve top ranked pages*, as counterpart measure of recall. Both measures rely on a human driven ranking of results, which is correlated with the search engine one in the first case. The second measure evaluates in which extent the top-results are present in each search engine, for the same query.

6.2 Experimental Setup

In this section, we describe in details our experimental setup. First of all, we define the evaluation goals:

- When searching for open data, how does STODaP compares to other data-specific and general search engines?
- What is the usability degree of STODaP server?

6.2.1 Subjects

The aim of STODaP server is to facilitate access to open data to the general public. We consider that experts already have their own strategies and sources for finding adequate data. Thus, we do not require experience in open data. However, users must have some previous knowledge on internet navigation. Knowledge on basic data processing tools such as spreadsheet processors is also desired, so that subjects can at least imagine a potential use of data.

In our experiment, participants were students attending a class on the topic Information Retrieval at the Federal University of Rio de Janeiro, in Brazil. An entry-

Table 12 – STODaP evaluation - summary of subjects profile

Participants	
Age	
Internet knowledge (1 - Never Used; 5 - Always use)	
Open Data Experience (1 - Never heard about; 5 - I work often with open data)	
Use of data (1 - I've never used any data processing tool; 5 - I'm an expert in advanced data processing tools)	

questionnaire was filled by the participants, whose answers are summarized in Table 12. Participation was not mandatory and there was no reward for participants.

6.2.2 Tasks

By design, STODaP server is a tool for interlinking different Open Data Portals. Thus, in this evaluation we aim to assess the ability of gathering similar information from several ODPs, rather than finding specific datasets on the Web.

The evaluation tasks were selected based on: (i) topic relevance of datasets on the open data community, based on criteria defined by Open Data Index¹ (ii) the existence of search results on STODaP server. This restriction allows us only to make assertions about the performance of STODaP server on the topics covered by the system, which consists of large base of open data portals, as described in Section 5.4. Broader conclusion would require large scale evaluations, which are over the scope of this thesis. Defined tasks are:

- Find open datasets containing budget information from 5 different countries.
- Find open datasets containing procurement information in 3 different idioms.
- Find open datasets about Water Quality on 10 different rivers.

6.3 Procedure

The following procedure was driven during the evaluation process:

- Participants filled the entry-questionnaire (5 minutes);
- The main idea of the project was explained, followed by an explanation about (10 minutes)
- Participants were assigned numbers and asked to enter this number on the form.
- Three tasks were sequentially presented. Each one was demanded to be completed

¹ <http://index.okfn.org/>

either using STODaP server, the Exversion Data Search Engine² or conventional search engines (XU; MEASE, 2009). The combination between search method, task and ordering was randomly chosen for participants.

- For each task, the challenge was presented with the appropriate number of text fields for pasting the results links.
- The time taken for each task was automatically calculated. The search string used in STODaP server were also captured.
- An evaluation questionnaire was filled by the subjects, containing questions about usability and satisfaction.

6.4 Results

6.4.1 Validation

Each entry-questionnaire was analysed in order to determine if it is valid to our evaluation, in terms of internet experience.

The answers were also checked in order to confirm if the dataset links provided are really valid answers to the assigned task.

6.4.2 Analysis

- Task completion time for each task (and variance) (XU; MEASE, 2009)
- Task completion time for each search method (and variance) (XU; MEASE, 2009)
- Correlation between satisfaction and completion time for STODaP server (XU; MEASE, 2009)
- Correlation between results found in the different search methods (VAUGHAN, 2004)

6.5 Conclusions

² Available at <<https://www.exversion.com/search/>>

7 Conclusions

The open data promises are still far from being materialised. However, it cannot be denied that significant advances have been made in recent years. In this section, the conclusions of this thesis will be derived, based on the initial hypothesis and in other findings that were made during this work.

In the Introduction of this work, two hypothesis were posed, which, for the ease of reading, are reproduced here:

H1: Enhancing the organization of open data repositories leads to better use of open data;

H2: Increasing the level of data literacy on the society leads to better use of open data (which in turn motivates better publishing).

Regarding H1, in the previous chapter results shown that system developed helped people in finding the more datasets quicker than using traditional approaches.

Bibliography

ADAMS, T.; STRECK, D. R. Educação Popular e novas tecnologias. *Educação*, v. 33, n. 2, p. 119–127, 2010. Cited 2 times on pages 40 and 44.

ALLAHYARI, M. Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network. In: *Proc. of the 10th International Conference on Semantic Computing*. [S.l.: s.n.], 2016. ISBN 9781509006625. Cited on page 72.

ALVEAR, C. A. S. de. *Tecnologia e participação: sistemas de informação e a construção de propostas coletivas para movimentos sociais e processos de desenvolvimento local*. 299 p. Tese (Tese de Doutorado) — Universidade Federal do Rio de Janeiro, 2014. Cited on page 37.

ANGELETOU, S. Semantic Enrichment of Folksonomy Tagspaces. In: _____. *ISWC*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 889–894. ISBN 978-3-540-88564-1. Cited 3 times on pages 69, 70, and 71.

ANGELETOU, S.; SABOU, M.; MOTTA, E. Semantically enriching folksonomies with FLOR. *Workshop of Collective Semantics*, v. 351, p. 65–79, 2008. ISSN 16130073. Cited on page 70.

ATKINSON, P.; HAMMERSLEY, M. Ethnography and Participant Observation. In: DENZIN, N. K.; LINCOLN, Y. S. (Ed.). *Handbook of qualitative research*. Thousand Oaks: Sage, 1994. p. 248–260. Cited on page 59.

ATTARD, J.; ORLANDI, F.; AUER, S. Value Creation on Open Government Data. In: *Proc. of the 49th Hawaii International Conference on System Sciences*. Kauai: [s.n.], 2016. p. 10. Cited 2 times on pages 29 and 33.

ATTARD, J. et al. A Systematic Review of Open Government Data Initiatives. *Government Information Quarterly*, v. 32, n. 4, p. 399–418, 2015. ISSN 0740-624X. Disponível em: <<http://dx.doi.org/10.5281/zenodo.18592>>. Cited 2 times on pages 21 and 35.

AUER, S.; BRYL, V.; TRAMP, S. *Linked Open Data – Creating Knowledge Out of Interlinked Data*. Cham: Springer International Publishing, 2014. v. 8661. (Lecture Notes in Computer Science, v. 8661). ISBN 978-3-319-09845-6. Disponível em: <<http://link.springer.com/10.1007/978-3-319-09846-3>>. Cited on page 61.

BARATO, J. N. *Codification/decodification: an experimental investigation on the adult education theory of Paulo Freire*. Tese (Thesis) — San Diego State University, 1984. Disponível em: <<https://jarbas.wordpress.com/048-codificacaodecodificacao-em-paulo-freire/>>. Cited on page 43.

BARGH, M. S.; CHOENNI, S.; MEIJER, R. Meeting Open Data Halfway: On Semi-Open Data Paradigm. In: *Proc. of the 9th International Conference on Theory and Practice of Electronic Governance - ICEGOV*. Montevideo: [s.n.], 2016. Cited on page 24.

BEGHIN, N.; ZIGONI, C. *Measuring Open Data's Impact of Brazilian National and Sub-National Budget Transparency Websites and its Impacts on Peoples's rights*. Brasília, 2014. Cited on page 29.

BERNERS-LEE, T. Linked Data - Design Issues. *W3C Website*, 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Cited on page 33.

BERNERS-LEE, T. *5 Stars Open Data*. 2010. Disponível em: <http://5stardata.info/>. Cited 2 times on pages 23 and 31.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. *Scientific American*, v. 284, n. 5, p. 34–43, 2001. ISSN 0036-8733. Cited on page 34.

BHARGAVA, R. *Towards a Concept of "Popular Data"*. 2013. Disponível em: <https://datatherapy.org/2013/11/18/towards-a-concept-of-popular-data/>. Cited on page 40.

BHARGAVA, R.; IGNAZIO, C. D. Designing Tools and Activities for Data Literacy Learners. In: *I Data Literacy Workshop*. Oxford: [s.n.], 2015. Cited 2 times on pages 16 and 38.

BIRD, S.; LOPER, E.; KLEIN, E. *Natural Language Processing with Python*. [S.l.]: O'Reilly Media Inc., 2009. Cited on page 93.

BRANDÃO, C. R. *O que é o método Paulo Freire*. São Paulo: Brasiliense, 1985. Cited on page 43.

CAPLAN, R. et al. *Towards common methods for assessing open data: workshop report & draft framework*. New York, 2014. v. 31, 1–15 p. Cited on page 29.

CARLSON, J. et al. Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *Libraries Faculty and Staff Scholarship and Research*, v. 11, n. 2, p. 629–657, 2011. ISSN 1530-7131. Cited on page 38.

CATTUTO, C. et al. Semantic grounding of tag relatedness in social bookmarking systems. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [S.l.: s.n.], 2008. v. 5318 LNCS, p. 615–631. ISBN 3540885633. ISSN 03029743. Cited on page 71.

CHEMUDUGUNTA, C. et al. Modeling documents by combining semantic concepts with unsupervised statistical learning. In: *The Semantic Web - ISWC*. [S.l.: s.n.], 2008. p. 229–244. ISBN 3540885633. ISSN 03029743. Cited on page 72.

CHENG, J.; HU, X.; HEIDORN, P. B. New measures for the evaluation of interactive information retrieval systems: Normalized task completion time and normalized user effectiveness. In: *Proceedings of the ASIST Annual Meeting*. [S.l.: s.n.], 2010. v. 47, n. April 2016. ISBN 1450470114. ISSN 15508390. Cited on page 99.

CHIGNARD, S. *A Brief History of Open Data*. 2013. Disponível em: <http://www.paristechreview.com/2013/03/29/brief-history-open-data/>. Cited 2 times on pages 21 and 38.

CILIBRASI, R. L.; VITANYI, P. M. B. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, v. 19, n. 3, p. 370–383, 2007. ISSN 10414347. Cited on page 71.

COLPAERT, P. et al. The 5 stars of open data portals. In: *7th international conference on methodologies, technologies and tools enabling e-Government (MeTTeG)*. [S.l.: s.n.], 2013. p. 61–67. Cited 2 times on pages 73 and 83.

COLPAERT, P. et al. Quantifying the interoperability of open government datasets. *Computer*, v. 47, n. 10, p. 50–56, 2014. ISSN 00189162. Cited on page 69.

CORAZZA, S. M. *Tema Gerador: concepção e prática*. Ijuí: Editora Unijuí, 2003. Cited on page 43.

CYGANIAK, R.; MAALI, F.; PERISTERAS, V. Self-service linked government data with dcat and gridworks. In: PASCHKE, A.; HENZE, N.; PELLEGRINI, T. (Ed.). *I-SEMANTICS*. [S.l.]: ACM, 2010. ISBN 978-1-4503-0014-8. Cited 2 times on pages 16 and 83.

Data Revolution Group. *A World That Counts - Mobilising the Data Revolution for Sustainable Development*. [S.l.], 2014. 32 p. Disponível em: <<http://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf>>. Cited 3 times on pages 16, 33, and 37.

DAVIES, T. *Open data, democracy and public sector reform*. 1–47 p. Tese (Dissertation) — University of Oxford, 2010. Cited 2 times on pages 31 and 32.

DAVIES, T. Supporting open data use through active engagement. In: *Proceedings of the W3C Using Open Data Workshop*. Brussels: [s.n.], 2012. p. 1–5. Cited 2 times on pages 30 and 31.

DAVIES, T.; SHARIF, R. M.; ALONSO, J. M. *Open Data Barometer - Global Report - 2nd Edition*. The World Wide Web Foundation, 2015. 1–62 p. Disponível em: <<http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf>>. Cited 2 times on pages 25 and 33.

DAVIES, T. G.; BAWA, Z. A. (Ed.). *Community Informatics and Open Government Data*. Vol 8, no. [S.l.]: Journal of Community Informatics, 2012. Cited on page 35.

DING, L. et al. TWC LOGD: A portal for linked open government data ecosystems. *Journal of Web Semantics*, Elsevier B.V., v. 9, n. 3, p. 325–333, sep 2011. ISSN 15708268. Cited on page 72.

DING, L. et al. {TWC} LOGD: A portal for linked open government data ecosystems. *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 9, n. 3, p. 325–333, 2011. ISSN 1570-8268. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1570826811000382>>. Cited 2 times on pages 16 and 82.

EAVES, D. *The Three Laws of Open Government Data*. 2009. Disponível em: <<http://eaves.ca/2009/09/30/three-law-of-open-government-data/>>. Cited on page 24.

ERMILOV, I.; AUER, S.; STADLER, C. User-driven semantic mapping of tabular data. In: SABOU, M. et al. (Ed.). *I-SEMANTICS 2013*. ACM, 2013. p. 105–112. ISBN 978-1-4503-1972-0. Disponível em: <<http://doi.acm.org/10.1145/2506182.2506196>>. Cited 2 times on pages 16 and 82.

ERMILOV, I.; AUER, S.; STADLER, C. User-driven semantic mapping of tabular data. In: SABOU, M. et al. (Ed.). *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*. ACM, 2013. p. 105–112. ISBN 978-1-4503-1972-0. Disponível em: <<http://doi.acm.org/10.1145/2506182.2506196>>. Cited on page 72.

FALS-BORDA, O.; RAHMAN, M. A. *Action and knowledge: breaking the monopoly with participatory action-research*. [S.l.]: Appex Press, 1991. Cited on page 37.

FELLBAUM, C. *WordNet: An Electronic Lexical Database*. 1998. 423 p. Cited on page 77.

FERRARO, A. R.; KREIDLOW, D. Analfabetismo no Brasil: configuração e gênese das desigualdades regionais. *Educação e Realidade*, v. 29, n. 2, p. 179–200, 2004. Cited on page 41.

FERREIRA, S. d. L.; SANTOS, E. O. dos. Formação de professores e cibercultura: novas práticas curriculares na educação presencial e a distância. In: *IV ANPED-Sul Seminário de Pesquisa em Educação da Região Sul*. Florianópolis: UFSC, 2002. v. 1. Cited on page 40.

FIORETTI, M. *A proposal to promote Open Data from and for the schools*. 2011. Disponível em: <<http://mfioretti.com/2011/10/warsaw-open-data-and-education/>>. Cited on page 39.

FREIRE, P. *Educação e Mudança*. 12. ed. Rio de Janeiro: Paz e Terra, 1979. Cited on page 43.

FREIRE, P. *Pedagogia do Oprimido*. 11. ed. [S.l.]: Editora Paz e Terra, 1987. Cited on page 41.

FREIRE, P. *Pedagogy of the Oppressed*. 30. ed. New York: Continuum, 2005. ISBN 8521902433. Disponível em: <<http://www.infed.org/thinkers/et-freir.htm>>. Cited on page 43.

GADOTTI, M. *Paulo Freire: Uma Biobibliografia*. São Paulo: Cortez Editora/Instituto Paulo Freire, 1996. Cited on page 42.

GHISO, A. M. Sistematización: un pensar el hacer que se resiste a perder su autonomía. *Decisio*, v. 1, n. 28, p. 3–8, 2011. Cited on page 44.

GRANICKAS, K. *Understanding the impact of releasing and re-using open*. [S.l.], 2013. 1–29 p. Cited on page 29.

GREY, J.; BOUNEGRU, L.; CHAMBERS, L. *Data Journalism Handbook*. [S.l.]: OKFN, 2012. Cited on page 38.

GRUBBER, T. Ontology of Folksonomy: A Mash Up of Apples and Organges. *Int'l Journal on Semantic Web & Information Systems*, v. 3, n. 2, 2007. Disponível em: <<http://tomgruber.org/writing/ontology-of-folksonomy.htm>>. Cited 2 times on pages 66 and 67.

GURSTEIN, M. B. Open data: Empowering the empowered or effective data use for everyone? *First Monday*, v. 16, n. 2, p. 1–7, 2011. Cited 2 times on pages 26 and 33.

GURSTEIN, M. B. Why I'm Giving Up on the Digital Divide. *Journal Of Community Informatics*, v. 11, n. 1, 2015. Disponível em: <<http://ci-journal.net/index.php/ciej/article/view/1210/1139>>. Cited 2 times on pages 38 and 45.

HALPIN, H.; ROBU, V.; SHEPHERD, H. The complex dynamics of collaborative tagging. In: *International World Wide Web Conference*. [s.n.], 2007. p. 211–220. ISBN 9781595936547. Disponível em: <<http://portal.acm.org/citation.cfm?id=1242602>>. Cited on page 67.

HARISPE, S. et al. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, v. 30, n. 5, p. 740–742, 2014. Disponível em: <<http://bioinformatics.oxfordjournals.org/content/30/5/740.abstract>>. Cited on page 71.

HARISPE, S. et al. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, v. 30, n. 5, p. 740–742, 2014. Disponível em: <<http://bioinformatics.oxfordjournals.org/content/30/5/740.abstract>>. Cited on page 91.

HARISPE, S. et al. Semantic Similarity from Natural Language and Ontology Analysis. *Synthesis Lectures on Human Language Technologies*, v. 8, n. 1, p. 1–254, 2015. Disponível em: <<http://dx.doi.org/10.2200/S00639ED1V01Y201504HLT027>>. Cited 2 times on pages 71 and 76.

HERN, A. *New York taxi details can be extracted from anonymised data*. 2014. Disponível em: <<http://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn>>. Cited on page 30.

HUIJBOOM, N.; BROEK, T. V. D. Open data: an international comparison of strategies. *European Journal of ePractice*, v. 1, n. 12, p. 1–13, 2011. Cited 3 times on pages 20, 38, and 39.

JARA, O. Los desafíos de la educación popular. In: *Metodología de La Educación Popular*. La Habana: Asociación de Pedagogos de Cuba, 1998. Cited on page 43.

JETZEK, T.; AVITAL, M.; BJØRN-ANDERSEN, N. Generating Value from Open Government Data. *ICIS 2013 Proceedings*, p. 1–20, 2013. Disponível em: <<http://aisel.aisnet.org/icis2013/proceedings/GeneralISTopics/5>>. Cited on page 29.

KIM, H. L. et al. The State of the Art in Tag Ontologies: A Semantic Model for Tagging and Folksonomies. In: *Proc. Int'l Conf. on Dublin Core and Metadata Applications*. [s.n.], 2008. p. 128–137. ISBN 3940344494. ISSN 3940344494. Disponível em: <<http://dcpapers.dublincore.org/ojs/pubs/article/view/925>>. Cited on page 67.

KIM, H. L. et al. Integrating tagging into the web of data: Overview and combination of existing tag ontologies. *Journal of Internet Technology*, v. 12, n. 4, p. 561–572, 2011. ISSN 16079264. Cited on page 67.

KNERR, T. *Tagging ontology-towards a common ontology for folksonomies*. [S.l.], 2006. 3–8 p. Disponível em: <<https://tagont.googlecode.com/files/TagOntPaper.pdf>>. Cited 2 times on pages 7 and 67.

KRÖTZSCH, M. et al. Semantic wikipedia. *Journal of Web Semantics*, December 2007. Disponível em: <http://www.aifb.uni-karlsruhe.de/Publicationen/showPublikation_english?publ_id=1551>. Cited on page 94.

LANIADO, D.; MIKA, P. Making sense of Twitter. In: *ISWC*. [S.l.: s.n.], 2010. Cited on page 69.

LAWLER, R. et al. Open Reconcile: A practical open-sourced ontology-driven webservice. *Proceedings of the 2012 IEEE 16th International Enterprise Distributed Object Computing Conference Workshops, EDOCW 2012*, n. 1, p. 124–131, 2012. Disponível em: <http://ieeexplore.ieee.org/xpls/abs/_all.jsp?arnumber=6406217http://ieeexplore.ieee.org/xpls/abs/_all.jsp?arnumber=6406217>. Cited on page 70.

LIMPENS, F.; GANDON, F.; BUFFA, M. A complete life-cycle for the semantic enrichment of folksonomies. In: GUILLET, F. et al. (Ed.). *Advances In Knowledge Discovery and Management*. [S.l.]: Springer Berlin Heidelberg, 2013. p. 127–150. ISBN 9783642358548. Cited 4 times on pages 7, 66, 71, and 72.

LOHMANN, S.; DÍAZ, P.; AEDO, I. MUTO: the modular unified tagging ontology. In: *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics '11*. [s.n.], 2011. p. 95–104. ISBN 9781450306218. ISSN <null>. Disponível em: <<http://dl.acm.org/citation.cfm?id=2063531>>. Cited on page 67.

MARCHETTI, A.; ROSELLA, M. SemKey : A Semantic Collaborative Tagging System. In: *Proceedings of the 16th international conference on World Wide Web - WWW '07*. [S.l.: s.n.], 2007. v. 7, p. 8–12. Cited 4 times on pages 66, 68, 70, and 79.

MELO, G. D. Lexvo . org : Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web Journal*, v. 7, p. 1–5, 2013. Cited on page 77.

MIKA, P. Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, v. 5, n. 1, p. 5–15, 2007. ISSN 15708268. Cited on page 66.

MKUDE, C. G.; PÉREZ-ESPÉS, C.; WIMMER, M. a. Participatory budgeting: A framework to analyze the value-add of citizen participation. *Proceedings of the Annual Hawaii International Conference on System Sciences*, p. 2054–2062, 2014. ISSN 15301605. Cited on page 26.

MURRAY-RUST, P. Open Data in Science. *Serials Review*, v. 34, p. 52–64, 2008. ISSN 00987913. Cited on page 20.

NEWMAN, R. *Tag ontology design*. 2005. Disponível em: <<http://www.holygoat.co.uk/projects/tags/>>. Cited on page 67.

NUÑEZ, C. Educar para transformar, transformar para educar. In: *Metodología de La Educación Popular*. La Habana: Asociación de Pedagogos de Cuba, 1998. Cited on page 43.

OBAMA, B. *Memorandum for the Heads of Executive Departments and Agencies*. The White House, 2009. Disponível em: <https://www.whitehouse.gov/the{_}press{_}office/TransparencyandOpenGove>. Cited on page 21.

OCHOA, X.; DUVAL, E. Quality Metrics for Learning Object Metadata. In: *World Conference on Educational Multimedia, Hypermedia and Telecommunications*. [S.l.: s.n.], 2006. ISBN 1-880094-60-6. Cited on page 69.

Open Knowledge Foundation. *Open Data Handbook*. [S.l.]: OKFN, 2015. Cited on page 20.

OPENSPEEDING. *Budget Data Package*. [S.l.], 2014. Disponível em: <<https://github.com/openspeeding/budget-data-package>>. Cited 2 times on pages 27 and 28.

PARYCEK, P.; SCHÖLLHAMMER, R.; SCHOSSBÖCK, J. “Each in Their Own Garden”: Obstacles for the Implementation of Open Government in the Public Sector of the German-speaking Region. In: *Proc. of the 9th International Conference on Theory and Practice of Electronic Governance - ICEGOV*. Montevideo: [s.n.], 2016. Cited on page 30.

PASSANT, A. Linked Data tagging with LODr. In: *Semantic Web Challenge (International Semantic Web Conference)2*. [S.l.: s.n.], 2008. p. 1–8. Cited 3 times on pages 67, 77, and 79.

REICHE, K. J.; HOFIG, E. Implementation of metadata quality metrics and application on public government data. In: *Proceedings - International Computer Software and Applications Conference*. [S.l.: s.n.], 2013. p. 236–241. ISBN 9780769549873. ISSN 07303157. Cited on page 69.

RENZIO, P. D.; WEHNER, J. *The Impacts Openness: A Review of the Evidence*. [S.l.], 2015. 35 p. Cited on page 29.

ROBU, V.; HALPIN, H.; SHEPHERD, H. Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web*, v. 3, n. 4, 2009. ISSN 15591131. Disponível em: <<http://eprints.soton.ac.uk/268192/>>. Cited on page 95.

ROSEIRA, C. Exploring the Barriers in the Commercial Use of Open Government Data. In: *Proc. of the 9th International Conference on Theory and Practice of Electronic Governance - ICEGOV*. Montevideo: [s.n.], 2016. Cited on page 30.

SANTOS, B. d. S. *A Gramática do Tempo - Para uma Nova Cultura Política - Col. Para um Novo Senso Comum - Vol. 4*. [S.l.]: Cortez, 2006. Cited on page 45.

SCHILD, M. Information Literacy, Statistical Literacy and Data Literacy. *IASSIST Quarterly Summer/Fall*, v. 28, n. 2/3, p. 6–11, 2004. Disponível em: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.144.6309>>. Cited 2 times on pages 16 and 39.

School of Data. *School of Data Report 2014*. 2014. Disponível em: <<http://2014.schoolofdata.org/>>. Cited on page 39.

SCHUGURENSKY, D. *Paulo Freire*. London: Bloomsbury Publishing, 2014. Cited on page 42.

SCHULER, D.; NAMIOKA, A. *Participatory Design: Principles and Practices*. [S.l.]: Lawrence Erlbaum Associates, 1993. Cited on page 37.

SIGURBJÖRNSSON, B.; ZWOL, R. V. Flickr tag recommendation based on collective knowledge. *Proceedings of the 17th international conference on World Wide Web - WWW '08*, v. 6, p. 327–336, 2008. ISSN 08963207. Disponível em: <<http://dl.acm.org/citation.cfm?id=1367542>>. Cited on page 95.

SMITH, B. Ontology. *Blackwell Guide to the Philosophy of Computing and Information*, n. 1964, p. 155–166, 2003. ISSN 1943-4723. Cited on page 34.

SPECIA, L. et al. Integrating Folksonomies with the Semantic Web. *Lecture Notes in Computer Science - The Semantic Web: Research and Applications*, v. 4519, n. September 2006, p. 624–639, 2007. ISSN 0302-9743. Disponível em: <<http://www.springerlink.com/content/413285327hj53234/>>. Cited 2 times on pages 70 and 71.

TAUBERER, J. *Open Government Data: The Book (2nd Edition)*. [S.l.]: Author's Edition, 2014. Cited 3 times on pages 21, 23, and 35.

TRILLO, R. et al. Discovering the Semantics of User Keywords. *Journal of Universal Computer Science*, v. 13, n. 12, p. 1908–1935, 2007. ISSN 0948-695X. Cited on page 71.

TYGEL, A. F. et al. “How much?” Is Not Enough An Analysis of Open Budget Initiatives. In: *Proc. of the 9th International Conference on Theory and Practice of Electronic Governance - ICEGOV*. Montevideo: [s.n.], 2016. p. 10. Cited 2 times on pages 18 and 26.

TYGEL, A. F. et al. Towards Cleaning-up Open Data Portals: A Metadata Reconciliation Approach. In: *Proc. of the 10th International Conference on Semantic Computing*. Laguna Hills, California: [s.n.], 2016. p. 8. Disponível em: <<http://arxiv.org/abs/1510.04501>>. Cited on page 18.

TYGEL, A. F.; CAMPOS, M. L. M.; ALVEAR, C. A. S. de. Teaching Open Data for Social Movements - a Research Methodology. *Journal of Community Informatics*, v. 11, n. 3, 2015. Disponível em: <<http://ci-journal.net/index.php/ciej/article/view/1220/1165>>. Cited 5 times on pages 18, 41, 47, 51, and 57.

TYGEL, A. F.; KIRSCH, R. Contributions of Paulo Freire for a critical data literacy. In: *I Data Literacy Workshop*. Oxford: [s.n.], 2015. p. 5. Cited 6 times on pages 18, 41, 47, 120, 121, and 122.

UMBRICH, J.; NEUMAIER, S.; POLLERES, A. Quality assessment & evolution of Open Data portals. In: *The International Conference on Open and Big Data*. [S.l.: s.n.], 2015. Cited 4 times on pages 69, 75, 77, and 83.

VAFOPOULOS, M. et al. Insights in global public spending. In: *Proceedings of the 9th International Conference on Semantic Systems - I-SEMANTICS '13*. [s.n.], 2013. p. 135–139. ISBN 9781450319720. Disponível em: <<http://dl.acm.org/citation.cfm?id=2506182.2506201>>. Cited on page 26.

VAHEY, P.; YARNALL, L.; PATTON, C. Mathematizing middle school: Results from a cross-disciplinary study of data literacy. In: *American Educational Research Association Annual Conference*. [S.l.: s.n.], 2006. p. 1–15. Cited on page 39.

Van Hooland, S. et al. Evaluating the success of vocabulary reconciliation for cultural heritage collections. *Journal of the American Society for Information Science and Technology*, v. 64, n. 3, p. 464–479, 2013. ISSN 15322882. Cited on page 70.

VAUGHAN, L. New measurements for search engine evaluation proposed and tested. *Information Processing and Management*, v. 40, n. 4, p. 677–691, 2004. ISSN 03064573. Cited 2 times on pages 100 and 102.

VLASOV, V.; PARKHIMOVICH, O. Development of the Open Budget Format. In: *Proceedings of the 16th conference of fruct association association*. Oulu: [s.n.], 2014. p. 129–136. Cited on page 27.

VLEUGELS, R. Overview of all FOI laws. *Fringe Special*, p. 1–28, 2012. Cited 2 times on pages 15 and 26.

WAAL, S. van der et al. Lifting Open Data Portals to the Data Web. In: AUER, S.; BRYL, V.; TRAMP, S. (Ed.). *Linked Open Data – Creating Knowledge Out of Interlinked Data*. [S.l.]: Springer, 2014. cap. 9. Cited on page 72.

WEI, B. et al. A survey of faceted search. *Journal of Web Engineering*, v. 12, n. 1&2, p. 41–64, 2013. ISSN 1540-9589. Cited on page 99.

WOLFF, A.; KORTUEM, G.; CAVERO, J. Urban Data in the primary classroom: bringing data literacy to the UK curriculum. In: *I Data Literacy Workshop*. Oxford: [s.n.], 2015. Disponível em: <<http://oro.open.ac.uk/43855/1/webSci-CR.pdf>>. Cited on page 39.

WORTHY, B. *David Cameron's Transparency Revolution? The Impact of Open Data in the UK*. London, 2013. Cited on page 26.

WU, X.; ZHANG, L.; YU, Y. Exploring social annotations for the semantic web. *Proceedings of the 15th international conference on World Wide Web - WWW '06*, p. 417, 2006. Disponível em: <[http://doi.acm.org/10.1145/1135777.1135839\\$\\delimiter\"026E30F\\$nhhttp://portal.acm.org/citation.cfm?id=1135777.1135839\\$\\delimiter\"026E30F\\$nhhttp://portal.acm.org/citation.cfm?doid=1135777.1135839\\$\\delimiter\"026E30F\\$nhhttp://dl.acm.org/citation.cfm?id=1135777.1135839](http://doi.acm.org/10.1145/1135777.1135839$\\delimiter\)>. Cited on page 67.

XU, Y.; MEASE, D. Evaluating Web Search Using Task Completion Time. In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. [S.l.: s.n.], 2009. p. 676–677. ISBN 978-1-60558-483-6. Cited 2 times on pages 99 and 102.

ZUIDERWIJK, A.; JANSSEN, M. The Negative Effects of Open Government Data - Investigating the Dark Side of Open Data. In: *Proceedings of the 15th Annual International Conference on Digital Government Research*. Aguascalientes, Mexico: [s.n.], 2014. p. 147–152. ISBN 978-1-4503-2901-9. Disponível em: <<http://doi.acm.org/10.1145/2612733.2612761>>. Cited 3 times on pages 26, 30, and 32.

ZUIDERWIJK, A. et al. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, v. 10, n. 2, p. 156–172, 2012. Cited 4 times on pages [16](#), [30](#), [31](#), and [33](#).

Appendix

APPENDIX A – List of Publications

A.1 Peer-reviewed conferences

- TYGEL, A. F.; AUER, S.; DEBATTISTA, J., ORLANDI, F.; CAMPOS, M. L. M. . Towards Cleaning-up Open Data Portals: A Metadata Reconciliation Approach. To be presented at the 10th International Conference on Semantic Computing, Laguna Hills, California. February 3-5 2016.
- TYGEL, A. F.; ATTARD, J.; ORLANDI, F.; CAMPOS, M. L. M. ; AUER, S. . "How much?" Is Not Enough - An Analysis of Open Budget Initiatives. To be presented at ICEGOV 2016, Montevideo, March 1-3 2016.
- TYGEL, A. F. ; KIRSCH, R. . Contributions of Paulo Freire for a Critical Data Literacy. In: Data Literacy Workshop, 2015, Oxford. Proceedings of the Data Literacy Workshop, 2015.
- CARVALHO, L. ; RODRIGUES, F. ; FERREIRA, R. ; BRAGA, P. ; TYGEL, A. F. ; ALVEAR, C. A. S. ; PRIMO, R. . Software Livre e Metodologias Participativas - ensino e extensão em uma disciplina da Engenharia. In: Encontro Nacional de Engenharia e Desenvolvimento Social - ENEDS, 2015, Salvador. Anais do XII ENEDS, 2015.

A.2 Peer-reviewed journals

- TYGEL, A. F. ; LUIZA MACHADO CAMPOS, MARIA ; ALVEAR, C. A. S. . Teaching Open Data for Social Movements: a Research Strategy. Journal of Community Informatics, v. 11, p. 1, 2015.
- TYGEL, ALAN ; GONÇALVES, LEONARDO GONÇALVES ; SANTOS, MAYARA ; MARQUES, GABRIEL ; LUIZA MACHADO CAMPOS, MARIA . Informação para Ação: Desenvolvimento de um Portal de Dados Abertos Sobre Agrotóxicos. Revista Tecnologia e Sociedade, v. 11, p. 99-119, 2015.

A.3 Book chapters

- TYGEL, A. F. ; Tecnologias da Informação e Comunicação e Movimentos Sociais: o Caso da Cooperativa EITA. In: Felipe Addor e Flávio Chedid. (Org.). Tecnologia, Participação e Território - Reflexões a partir da Prática Extensionista. 1ed. Rio de Janeiro: Editora UFRJ / Faperj, 2015, v. 3, p. 259-292.

APPENDIX B – Results of Open Data Research

Table 13 – Motivations, Impediments and Improvements indicated in answers to Question 4.

Question 4: Why have you attended to the course? Why do you think open data is important?			
#	Motivations	Impediments	Improvements
4.1	Work with data and link different information to create arguments	There is a mismatch between amount of data released and the capacity of social movements to analyse it	Make investments in education for open data use
4.2	Be able to work with data driven journalism	There are many barriers to access information	Promote publicity about existence of data
4.3	Use data to denounce injustices	Open Data is unknown for most social movements	Improve knowledge about how to search for data
4.4	Data can give basis to stimulate new claims	There is no full transparency in government actions	Enable access to information, without discrimination
4.5	Translate data into information for readers	Most of the people have little informatics ability	
4.6	Produce data in juridical research		
4.7	Open data can stimulate analysis		
4.8	Open data can stimulate new data		
4.9	Validate/legitimate arguments in communication with data		
4.10	Use data to understand the capitalist society		
4.11	Understand the resistances against oppression with data		
4.12	Fight corruption using spending data		
4.13	Make better use of information, a central point in class conflicts		
4.14	Unveil data manipulation		

Source: Tygel e Kirsch (2015)

Table 14 – Impediments pointed in answers to Question 8.

Question 8: What is the main impediment perceived by using data?	
#	Impediments
8.1	The lack of knowledge about data production process makes interpretation difficult
8.2	It is hard to understand data connection and linking possibilities
8.3	Finding data in the web is hard /Open data portals are complicated
8.4	Access to data outside the web is hard / FoIA application is complicated
8.5	Data organization is confused
8.6	Data formats does not help its use
8.7	The state presents data through different platforms which increase the need for training
8.8	The need of specific software tools makes data usage harder
8.9	Some important data is concealed
8.10	Most data is outdated
8.11	The querying interfaces present too much information
8.12	Access to raw data is hard
8.13	Government agencies do not follow common data standards
8.14	"Data interpretation is difficult
8.15	Linking data from different sources is difficult without appropriate tools and metadata

Source: [Tygel e Kirsch \(2015\)](#)

Table 15 – Improvements indicated in answers to Question 9.

Question 9: How do you imagine that the use of data could be improved?	
#	Improvements
9.1	Provide user-friendly interfaces
9.2	Provide education on statistics/mathematics
9.3	Standardize open government data
9.4	Provide user-friendly language (avoid technical terms)
9.5	Provide wider training possibilities
9.6	Promote more advertising of open government data initiatives
9.7	Promote more advertising of social movements open data initiatives
9.8	Foster more research on open data and social movements
9.9	Improve data search engines
9.10	Increase the offer of open data sources
9.11	Avoid the need of intermediaries for data interpretation
9.12	Improve open data portals

Source: [Tygel e Kirsch \(2015\)](#)