

Universidade Federal do Rio de Janeiro
Instituto de Matemática & Instituto Tércio Pacitti de Aplicações e
Pesquisas Computacionais
Programa de Pós-Graduação em Informática

Alan Freihof Tygel

**Semantic Tags for Open Data Portals:
Metadata Enhancements for Searchable Open
Data**

Rio de Janeiro

2016

Alan Freihof Tygel

Semantic Tags for Open Data Portals: Metadata Enhancements for Searchable Open Data

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.

Supervisor: Maria Luiza M. Campos, PhD, UFRJ
Co-supervisor: Sören Auer, PhD, University of Bonn

Rio de Janeiro

2016

CIP - Catalogação na Publicação

T979s Tygel, Alan Freihof
Semantic Tags for Open Data Portals: Metadata
Enhancements for Searchable Open Data / Alan
Freihof Tygel. -- Rio de Janeiro, 2016.
161 f.

Orientadora: Maria Luiza Machado Campos.
Coorientador: Sören Auer.
Tese (doutorado) - Universidade Federal do Rio
de Janeiro, Instituto Tércio Pacitti de Aplicações
e Pesquisas Computacionais, Programa de Pós
Graduação em informática, 2016.

1. Open Data. 2. Semantic Web. 3. Metadata. 4.
Data Literacy. I. Campos, Maria Luiza Machado,
orient. II. Auer, Sören, coorient. III. Título.

Alan Freihof Tygel

Semantic Tags for Open Data Portals: Metadata Enhancements for Searchable Open Data

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.

Aprovada em 21 de julho de 2016.

Maria Luiza M. Campos, PhD, UFRJ
Orientadora

Sören Auer, PhD, University of Bonn
Co-orientador

Bernadette F. Lóscio, D.Sc., UFPE
Examinadora

Maria C. Cavalcanti, D.Sc., IME/RJ
Examinadora

Jonice de O. Sampaio, D.Sc., UFRJ
Examinadora

Marcelo Firpo de S. Porto, PhD,
Fiocruz
Examinador

Marcos Borges, Ph.D, UFRJ
Examinador

Acknowledgements

Durante a elaboração deste trabalho, diversas pessoas participaram direta ou indiretamente. Seria impossível citar todas elas, mas farei um esforço de pelo menos incluir quem irá se alegrar em ter seu nome registrado aqui. During the preparation of this work, several people participated directly or indirectly. It would be impossible to cite everyone, but I'll drive an effort to mention at least those who would be happy to see their names here. Während die Zubereitung von dieser Arbeit, viele Leute haben daran teilgenohmen, in eine direkte oder indirekte Art. Es währe unmöglich alle zu zitieren, aber ich versuche mindestens zu nennen die, die sich freuen werden, ihre Namen hier zu sehen.

Primeiramente, fora Temer! First of all, Temer out, illegitimate president of Brazil! Erstens, Raus mit Temer, illegitimer brasilianische Präsident!

Em segundo lugar, não por nenhuma contribuição especial à tese em si, mas por me lembrar que os agradecimentos são a parte mais lida da tese, e portanto merecem ser escritos com cuidado e carinho: obrigado, Flávio Chedid, por isso (e por tudo mais!).

Agora sim, pelas contribuições na tese e na vida, e por me mostrar que é possível construir uma outra informática na universidade, que é possível quebrar os paradigmas da sala de aula também na Engenharia, que é possível encantar os estudantes, mesmo aqueles que pensam diferente, que é possível (e tão necessário) debater de forma honesta em sala, sem impor e sem se fingir de neutro. Celso Alvear, meu amigo, obrigado por estar comigo até aqui, e certamente daqui pra frente!

Quando decidi mudar de ares (e áreas), ao sair dos processamentos sinais, foi no Núcleo de Solidariedade Técnica (Soltec) que me senti acolhido, e onde aprendi que a Engenharia pode ser revolucionária quando colocada a serviço do povo. Obrigado aos grandes mestres Sidney Lianza e Antônio Cláudio, eu aos colegas Felipe Addor, Fernanda Santos, Vicente Neponucemo, Marília Gonçalves, Amanda Azevedo, Renata Melo, Camile Perissé, Ricardo Mello, Jair Nastalino, Regina Carvalho, Walter Suemitsu, Sandra Mayrink, e não menos Soltec, o grande amigo Marcelo Ribeiro.

Um sonho que se sonha junto é realidade – muito obrigado às companheiras e companheiros da Cooperativa EITA por acreditarem que é possível colocar as tecnologias da informação nas mãos dos movimentos sociais: Rosana Kirsch, Fernanda Nagem, Daniel Tygel, Bráulio Bhavamitra, André Luís, Pedro Jatobá, Vinícius Brand e Bernardo Vaz.

Uma enorme parte da inspiração e motivação para este trabalho veio das companheiras e companheiros de luta pela reforma agrária e pela agroecologia. Se não for por isso, nada faz sentido: Nívia Silva, Jake Pivatto, Fran Paula, Karen Friedrich, Fernando

Carneiro, Raquel Rigotto, Lia Giraldo, Marcelo Firpo, Andre Burigo (Deco), Leonardo Melgarejo, Fábio Miranda, Íris Pacheco, Carla Bueno, e tantas, e tantos muitos outros!

Durante os 4 anos desta tese, muitas crianças nasceram. Nelas deposito a esperança da continuidade da luta por um mundo melhor: Ernesto (de Nanda e Rafa), Iara (de Jojo e Victor), Pedro (de Aline), Manu (de Camila e Juan), Bruno (de Mara e Balso), Sofia (de Celso e Karen) e Clara (de Bernardo e Nana).

To my colleagues at Uni-Bonn and Fraunhofer, who shared technical discussions, beers, barbecues and pizzas with me: Judie, Fabrizio, Simon, Jeremy, Sidra, Farah, Denis, los muchachos Diego and Irlan, who made Bonn a bit more latin, Harsh – the EIS greatest ping pong player, Mohamed, Michael-el-russo, Matthew, Nicklas, Lavdim, Cristoph, Steffen, Farethin, Jaana, Najmeh, Kemele, Allen, Gökhan, Fathoni, Elisa, Christiane, Tiansi, Darya and Saeedeh.

And a special thanks for Prof. Sören Auer, for the warm welcome and for giving all the conditions for the development of this work.

2015 war ein ganz besonders Jahr in meine Leben. Für das erste mal, habe ein Jahr im Ausland gewohnt, und sogar in Deutschland. Glücklicherweise, konnte ich schon ein bisschen Deutsch sprechen, und damit konnte ich eine tiefere Lebenserfahrung habe. Dieses Jahr in Deutschland war absolut nicht nur über Studien. Zusammen mit die Doro48 WG - Stephi, Georg, Christian, Doro und Arne - und die Liebe erweitert WG: Miri, Kirsten, Soija, Nadine und Malte, habe ich ganz viel gekocht, Pilz gesammelt, gebummelt, Fahrrad gefahren, gereist, aber hauptsächlich ganz viel über Deutschland gelernt. Und Deutschland, im Moment, heißt auch den ganzen Welt. Heute darf ich zweifellos sagen, dass Bonn auch meinen Stadt ist.

Auch durch meine deutsche-mit-brasilianischen-Herz Freunde Mario, Manu und Ana, und die Brasilianer-die-Deutschland-ganz-gut-verstehen Tainã, Camila, Ana, Tico, konnte ich Deutschland (und Brasilien!) besser nachvollziehen.

Aos meus pais - Ruth, David, William, Monica - que me deram desde sempre as condições e apoio para fazer o que eu quisesse.

À minha companheira de vida, cozinha e sonhos: Uschi Silva, fonte de carinho e trocas fundamentais para ser quem eu sou.

Aos colegas do PPGI: Kelli Cordeiro, pelas conversas sobre as teses e a vida, Bruno Nascimento, Fabrício Firmino, Herli Menezes, Angélica Dias, Tiago Marino, Sérgio Serra, à ajuda fundamental dos servidores Aníbal e Adriana, e às/aos estudantes que tive o prazer de (des)orientar: Karen Teixeira, Mayara Santos, Leonardo Gonçalves e Raphaela Nunes.

À banca: Marcos Borges, Maria Cavalcanti (Yoko), Marcelo Firpo, desde a qualificação, e Jonice Oliveira e Bernardete Lóscio, agradeço pelas valorosas contribuições.

Ainda que não tenha sido possível incorporar tudo na versão, os comentários certamente contribuíram para o debate e trabalhos posteriores.

Finalmente, o agradecimento mais importante: Professora, Orientadora e amiga Maria Luiza. Quando terminei o mestrado, e comecei meu trabalho no Soltec em uma área distinta daquela que vinha trabalhando, tomei duas decisões: (i) se fosse fazer o doutorado, seria na área de Informática ou Engenharia de Computação, e não em nenhuma área menos técnica; (ii) só faria o doutorado após encontrar um/a orientador/a em que tivesse plena confiança de que iria embarcar nesta aventura junto comigo.

Logo nas primeiras conversas, desconfiei seriamente de que Luiza seria esta pessoa. Longe da arrogância infelizmente tão comum aos professores-doutores da nossa Universidade, ela reconheceu: não sou especialista nas áreas de desenvolvimento participativo e movimentos sociais, mas topo aprendermos juntos. E foi isso o que de fato aconteceu.

Nunca tive na vida uma relação tão horizontal com nenhum/a professor/a ou aluno/a como foi a nossa interação durante estes pouco mais de 4 anos. O sentido da dialogicidade, que tanto teorizamos, foi colocado em prática da forma mais completa que já experimentei. Nossas convergências e divergências políticas (e culinárias!), e o que conseguimos extrair disso me dá a certeza de que os diferentes olhares são profundamente necessários na construção do novo. Mas não adianta apenas ser diferente: tem que existir a capacidade de síntese.

Se um dia eu resolver me tornar professor universitário, tenho sua atuação como parâmetro. Não posso me tornar professor se não for para me dedicar de corpo e alma; se não for para olhar alunos/as não como seres sem luz, mas como pessoas que têm sua história, sua vida e sua condição social para além das notas das provas; se não for para lutar contra o conservadorismo, e a estagnação; e se não for para entender a informática enquanto meio de transformação social, nunca um fim em si mesma. Quando me achar em condições de contribuir para a Universidade como você contribui, aí sim, me sinto digno de ocupar esse lugar.

Luiza: durante a defesa você disse: “Se a tese não saiu como você gostaria, a culpa é toda minha.” E eu respondo: esta tese só saiu desta forma, e tão elogiada por uma banca tão heterogênea, por sua sensibilidade social, rigor científico, firmeza, humildade e capacidade de dialogar. Essa é a sua culpa.

Obrigado!

Preface

While looking at this (almost) finished thesis, it is impossible to avoid thinking about the initial motivations that took me to start this PhD. My several inquietudes regarding the role of Computer Science in our severely unequal society moved me to write a PhD project focused on how informatics can (positively) change the society. My experience on working with social movements made me believe that they are in the front line of the struggle against a system that will structurally deny forever the basic human rights for many human beings. These people – landless persons willing to have a piece of land in such an enormous country like Brazil; original populations, willing to maintain their culture in a piece of living forest; or homeless people facing the evil real estate speculation in our big cities, which leaves several void houses while people have to live in poor conditions – inspired me not only to enter this PhD, but to other several actions regarding support for their struggle.

Grassroots movements who organise these people have a tremendous power of transformation, mainly because their fight for social justice is the fair. In order to push these movements forward, several ingredients are necessary: hope, love, dedication, solidarity, and ... technology.

If the final result of this work did not generate concrete technological artefacts to support social movements, I'm sure that many intermediate results did, and many future works will do.

While looking at this (almost) finished thesis, it is also impossible to avoid thinking about the several places where I was in the last four years. On the first interview, I was talking from Salvador, in Bahia. The courses were all attended in Rio Janeiro, and the Data Literacy courses were also given in Porto Alegre and Vitória, besides Rio. The main contributions resulting from this thesis were discussed in Bonn. Several important decision were taken in a kitchen in Trento. And now, these last words are being written in Recife. Hopefully this thesis could get a little taste from all these wonderful places.

The decision of writing the text in English was not easy. Unfortunately, it turns the chances of reading and reusing in Brazil lower, since English speakers in our country are estimated in 5% of the population. However, I must recognize that much of this thesis was written over the work of researchers of several nationalities, and I didn't have to learn Chinese, Russian nor Dutch to read them. Thus, it is also fair that this piece comes out in English, and has the possibility of travelling even more than I did during its writing.

I wish a pleasant reading, and I also wish the future PhD Computer Science thesis can be more transdisciplinary than this was.

*“Deu meia noite, a lua faz o claro
Eu assubo nos aro, vou brincar no vento leste
A aranha tece puxando o fio da teia
A ciência da abeia, da aranha e a minha
Muita gente desconhece
Muita gente desconhece, olará, viu?
Muita gente desconhece”*

Na Asa do Vento, João do Vale e Luiz Vieira, 1981

Resumo

TYGEL, Alan Freihof. **Semantic Tags for Open Data Portals**: metadata enhancements for searchable open data. 2016. 161 f. Tese (Doutorado em Informática)-Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

A publicação massiva de dados em formatos abertos na Internet parece ser uma tendência irreversível. Espera-se com isso uma maior transparência das administrações públicas, que por sua vez alimenta a democracia com a população bem informada, e inibe o mau uso de recursos públicos através da possibilidade de uma inspeção pública dos gastos. Junto às grandes expectativas geradas pelas políticas de dados abertos, verifica-se também uma ampla gama de problemas. No decorrer desta pesquisa, dois problemas chamaram a atenção: (i) a falta de descritores adequados para conjuntos de dados abertos e (ii) as dificuldades do público em geral em lidar com os dados. Diversos estudos apontam que, 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 extremamente desiguais como o Brasil. Neste sentido, apresentamos nesta tese uma abordagem para alfabetização em dados, inspirada na pedagogia da educação popular e na pesquisa-ação participativa. A implementação desta abordagem como um trabalho de campo revelou que a má qualidade dos descritores dos conjuntos de dados abertos é um dos fatores que impedem o avanço dos dados abertos. Administradores dos portais de dados abertos utilizam diversos tipos de metadados para descrever seus conjuntos de dados, sendo as *tags* um dos mais importantes. Entretanto, o processo de atribuição de *tags* é sujeito a diversos problemas, como sinonímia, ambiguidade ou incoerência, entre outros. Face a estes problemas, nesta tese foi desenvolvida e implementada a abordagem de *Tags Semânticas para Portais de Dados Abertos* (STODaP, na sigla em inglês) – para limpeza, enriquecimento e conciliação de metadados em portais de dados abertos. A abordagem STODaP foi avaliada, e os resultados mostram que ela permitiu que os participantes do experimento encontrassem conjuntos de dados abertos mais rapidamente e de forma mais precisa do que utilizando outros métodos de busca. 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: Portal de Dados Abertos. Conciliação de Metadados. Enriquecimento Semântico. Alfabetização em Dados. Dados Abertos Interligados.

Abstract

TYGEL, Alan Freihof. **Semantic Tags for Open Data Portals**: metadata enhancements for searchable open data. 2016. 161 f. Tese (Doutorado em Informática)-Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

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 publish data government data through Open Data Portals (ODPs). Hence, it is expected a greater transparency of public administrations, which in turn feed democracy with a well informed population, and inhibits public resources misuse 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 adequate descriptors for open datasets, and (ii) the difficulties of the general public for dealing with open data. Thus, this thesis expects to bring a contribution for the field of open data by proposing an approach for these problems. Several studies attest that even if open 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. The application of this approach as a field study revealed that bad quality open datasets description is one of the factors hindering open data advance. ODP managers use several types of metadata to describe 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 developed and implemented the Semantic Tags for Open Data Portals approach, for metadata cleaning up, enriching and reconciliation in ODPs. The STODaP approach was evaluated, and results show that it enable participants to find open datasets faster and preciser than using other searching methods. It is expected that this thesis contributes 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 Portal. Metadata Reconciliaton. Semantic Lifting. Data Literacy. Linked Open Data.

List of Figures

Figure 1 – Data Spectrum as a definition of steps between open and closed data.	31
Figure 2 – Analysis framework open budget initiatives.	34
Figure 3 – Different uses of data, with process, summary and examples.	39
Figure 4 – Critical data literacy process.	54
Figure 5 – Trade-off between interpretation autonomy and software skills needed.	61
Figure 6 – Classification tree for open data engagement actions.	69
Figure 7 – Tagging Ontology.	74
Figure 8 – MOAT Ontology.	74
Figure 9 – MUTO Ontology.	75
Figure 10 – SRTag RDF schema.	80
Figure 11 – Re-use of tags inside an ODP.	82
Figure 12 – Average number of tags per dataset.	83
Figure 13 – Percentage of very similar tags in ODPs.	84
Figure 14 – Overview of the STODaP approach.	90
Figure 15 – Architecture of the STODaP approach.	91
Figure 16 – Relevant elements of an Open Data Portal.	92
Figure 17 – Local tag processor.	95
Figure 18 – Global metadata processor.	98
Figure 19 – Simplified schema of the STODaP vocabulary.	100
Figure 20 – Implementation architecture of the STODaP approach.	102
Figure 21 – STODaP welcome screen.	104
Figure 22 – STODaP Semantic Tags.	105
Figure 23 – Example of the <i>cadmium</i> Semantic Tag.	109
Figure 24 – STODaP faceted search.	110
Figure 25 – STODaP SPARQL endpoint.	111
Figure 26 – Screenshot of the Tag Manager plugin.	112
Figure 27 – Screenshot of the Semantig Tag plugin.	113
Figure 28 – Evaluation framework.	120
Figure 29 – Boxplot for TCT.	128
Figure 30 – Boxplot for TCT_{nb}	129
Figure 31 – Boxplot for TCT_c	130
Figure 32 – Precision analysis: Average.	131
Figure 33 – Precision analysis: Boxplot.	131

List of Tables

Table 1 – Decontextualized phrases used in the official literacy method, in Brazil.	49
Table 2 – Relation between Freire’s Literacy Method and data literacy.	52
Table 3 – Examples of data driven statements.	59
Table 4 – Open and closed analogies to help understand what open data is.	60
Table 5 – Examples of society driven databases.	62
Table 6 – Summary of the presentations of the open data course for social movements.	64
Table 7 – Questionnaire answered by course attendants.	65
Table 8 – Summary of data used in the experiment.	81
Table 9 – Expressiveness of tags.	86
Table 10 – Examples of tags in each step of the procedure.	97
Table 11 – Examples of groups in some ODPs.	99
Table 12 – Entry questionnaire.	119
Table 13 – Evaluation questionnaire.	121
Table 14 – Answers to the entry and evaluations questionnaires.	122
Table 15 – Task Completion Time of the pre-evaluation test, in seconds.	123
Table 16 – STODaP evaluation - summary of participants profile	125
Table 17 – Evaluation Results - TCT.	126
Table 18 – Evaluation Results - TCT_{nb}	127
Table 19 – Evaluation Results - TCT_c	127
Table 20 – STODaP evaluation - summary of subjective evaluation.	132
Table 21 – Correlation analysis of the results.	134
Table 22 – Motivations, Impediments and Improvements.	157
Table 23 – Impediments pointed in answers to Question 8.	158
Table 24 – Improvements indicated in answers to Question 9.	159

List of abbreviations and acronyms

CSO	Civil Society Organisation
CSV	Comma Separated Values
DCAT	Data Catalog Vocabulary
FOIA	Freedom of Information Act
ICT	Information and Communications Technology
LOD	Linked Open Data
NGO	Non-Governmental Organisation
MOAT	Meaning of a Tag
MUTO	Modular Unified Tagging Ontology
ODP	Open Data Portal
OGD	Open Government Data
OGP	Open Government Partnership
RDF	Resource Description Framework
SIOC	Semantically-Interlinked Online Communities
SKOS	Simple Knowledge Organization System
SPARQL	Protocol and RDF Query Language
STODaP	Semantic Tags for Open Data Portals
UN	United Nations
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	eXtensible Markup Language

Contents

1	INTRODUCTION	18
1.1	WHY OPEN DATA?	18
1.2	NOT ONLY ADVANTAGES – OPEN DATA PROBLEMS AND PERILS	19
1.3	HYPOTHESIS	21
1.4	OBJECTIVES	22
1.5	SOLUTION APPROACH	22
1.6	METHODOLOGY	23
1.7	STRUCTURE OF THE THESIS	24
2	OPEN DATA – AN OVERVIEW	25
2.1	WHY OPEN DATA?	25
2.2	HISTORICAL NOTES	27
2.3	DEFINITIONS	28
2.4	OPEN DATA LANDSCAPE	31
2.5	OPEN BUDGET DATA	33
2.6	EVALUATING OPEN DATA IMPACTS AND VALUE	35
2.7	PROBLEMS OF OPEN DATA	37
2.8	LINKED DATA TOWARDS SEMANTIC DESCRIPTION OF OPEN DATA	41
2.9	CONCLUSIONS	43
3	OPEN DATA RESEARCH THROUGH DATA LITERACY	44
3.1	AN OVERVIEW ON DATA LITERACY	45
3.1.1	Data Literacy and Popular Education	47
3.2	CONTRIBUTIONS OF PAULO FREIRE FOR A CRITICAL DATA LITERACY	48
3.2.1	Paulo Freire, Literacy and Popular Education	48
3.2.1.1	Investigation Stage	49
3.2.1.2	Thematisation Stage	50
3.2.1.3	Problematism Stage	50
3.2.1.4	Systematisation Stage	51
3.2.2	Parallels between Literacy Education and Data Literacy	51
3.2.3	A Freirean Inspired Critical Data Literacy	52
3.2.3.1	The Emancipatory Character of Data Literacy	53
3.2.3.2	Data Literacy Process	53
3.2.3.3	Data Literacy Stages	54
3.2.3.4	Definition	57
3.2.4	Conclusions	57

3.3	TEACHING OPEN DATA FOR SOCIAL MOVEMENTS: ACTION AND RESEARCH FOR OPEN DATA ENGAGEMENT	58
3.3.1	First Stage – Introduction	59
3.3.2	Second Stage – Data Sources	60
3.3.3	Third Stage – Tools	63
3.3.4	Fourth Stage – Final Work	63
3.4	OPEN DATA CLUES FROM THE FIELD	64
3.4.1	Questionnaire Based Analysis	65
3.4.2	Observation Based Analysis	66
3.4.3	Synthesis	68
3.5	CONCLUSIONS	70
4	SEMANTIC METADATA FOR OPEN DATA DESCRIPTION	71
4.1	SEMANTIC METADATA: A LITERATURE REVIEW	71
4.1.1	Introduction	71
4.1.2	Characterization of the Contribution	75
4.1.3	Metadata Assessment	76
4.1.4	Metadata Clean-Up	77
4.1.5	Metadata Reconciliation	78
4.1.6	Structure Emergence	79
4.1.7	Automatic Semantic Tagging	79
4.1.8	Semantic Lifting in ODPs	80
4.2	AN ANALYSIS OF METADATA IN ODPS	80
4.2.1	Local Metrics	82
4.2.1.1	Tag Reuse	82
4.2.1.2	Tags per Dataset	82
4.2.1.3	Tag Similarity	83
4.2.2	Global Metrics	84
4.2.2.1	Coincident tags between portals	84
4.2.2.2	Tag expressiveness	85
4.3	CONCLUSIONS	86
5	STODAP APPROACH	87
5.1	MOTIVATION	87
5.2	STODAP ARCHITECTURE	89
5.2.1	Architecture Overview	90
5.2.2	Open Data Portals	91
5.2.3	ODP Extensions	93
5.2.4	Local Processor	95
5.2.5	Global Processor	97

5.2.6	Semantic Metadata Repository	99
5.2.7	STODaP Vocabulary	100
5.2.8	Interfaces	101
5.3	IMPLEMENTATION	102
5.3.1	Semantic Tags Server	102
5.3.2	Interfaces	103
5.3.3	CKAN Plugins	103
5.3.4	Use and Maintenance of the STODaP server	106
5.4	QUANTITATIVE RESULTS	106
5.4.1	STODaP Server	106
5.4.2	Local Level	107
5.5	CONCLUSIONS	107
6	EVALUATION	114
6.1	OVERVIEW	114
6.2	LITERATURE REVIEW	115
6.3	EXPERIMENTAL SETUP	116
6.3.1	Goals	117
6.3.2	Comparative Assessment	117
6.3.3	Participants	118
6.3.4	Questions	118
6.3.5	Procedure	119
6.3.6	Metrics	120
6.3.7	Validation	121
6.4	PRE-EVALUATION	122
6.5	EVALUATION	125
6.5.1	Participants Profile	125
6.5.2	Task Completion Time Analysis	125
6.5.3	Precision Analysis	129
6.5.4	Subjective Evaluation	132
6.5.5	Correlation Analysis	133
6.6	DISCUSSION	134
6.7	CONCLUSIONS	135
7	CONCLUSIONS	137
7.1	CONTRIBUTIONS	137
7.1.1	Main Contribution	137
7.1.2	Other Contributions	138
7.2	LIMITATIONS AND DIFFICULTIES	139
7.3	FUTURE WORK	141

REFERENCES 143

APPENDIX 153

APPENDIX A – LIST OF PUBLICATIONS 154

A.1 PEER-REVIEWED CONFERENCES 154
A.2 PEER-REVIEWED JOURNALS 154
A.3 BOOK CHAPTERS 154
A.4 SPECIAL ISSUE CO-EDITOR 155

APPENDIX B – RESULTS OF OPEN DATA RESEARCH 156

1 Introduction

This thesis is essentially about how to assist people who want to search for and use open data. Given the growing importance of open data to the society, the topic 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 an effort to contribute to a multidisciplinary field that is heavily related to Computer Science, but far from being restricted to it.

In this introductory chapter, some motivations behind the topic of open data are briefly exposed, highlighting the problems that still hinder people to have access to desired open datasets. The hypothesis from where this thesis starts is posed, as well as the main and specific objectives that we aim to achieve with this work. We further introduce our solution approach, and explain the methodology used in order to develop it. Finally, the structure of the remainder of this text is described.

1.1 Why Open Data?

Current numbers about the open data scene leave no doubt about the central importance of this topic in contemporary society. The Open Data Index¹ 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 Monitor² counts 173 open data catalogues in the continent, which sums an amount of 1472 GB of data.

The movement towards opening datasets has its roots related to a series of access-to-information laws. According to the right2info.org³ platform, these are laws that “establish the right and procedures for the public to request and receive government-held information”. [Yannoukakou and Araka \(2014\)](#) drive a comprehensive description about the synergies between the right to information and open data movements. According to them, both movements can push the formulation of an universal approach on access of government 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.

¹ Available at <http://index.okfn.org/place/>.

² Available at <http://opendatamonitor.eu>.

³ Available at <http://right2info.org>.

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 have 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 OpenDataSoft⁴. The potential of changing the very basis of democratic processes took the United Nations (UN) to coin 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, p.4). Still according to the UN, these data are “creating unprecedented possibilities for informing and transforming society and protecting the environment” (Data Revolution Group, 2014, p.4).

A study by Huijboom and Broek (2011) comparing national open data strategies of five countries analysed their public policies programmes and key motivations behind publishing open data. Authors defined three main categories of motivations:

- Increase democratic control and political participation, i.e., open data could empower citizens on exercising their democratic rights;
- Foster service and product innovation, i.e., open data could generate new opportunities for innovation on the public and private sector. Under this topic, it has been recently stated that “Open data can help unlock U\$3 trillion to U\$5 trillion in economic value annually” (MANYIKA et al., 2013).
- Strengthen law enforcement, i.e., open data could enable citizen involvement and enable the development of security application.

While the big numbers and great expectations about open data may generate an enthusiastic hope that this movement will solve many problems of the society, there are also critical voices claiming that the promises are still far from being realised, and also that there are some hidden threats that should be alerted. The problems of open data are the subject of next section.

1.2 Not Only Advantages – Open Data Problems and Perils

In front of such a big hope regarding the benefits of opening data, several authors have also dedicated themselves to analyse the topic from a critical point of view (ZUIDERWIJK et al., 2012; ZUIDERWIJK; JANSSEN, 2014a; GURSTEIN, 2011; BATES, 2014; ROSEIRA, 2016; PARYCEK; SCHÖLLHAMMER; SCHOSSBÖCK, 2016; DAVIES; BAWA, 2012)⁵.

⁴ Available at <<https://www.opendatasoft.com>>.

⁵ Open data problems will be deeper analysed in Section 2.7. For the purposes of this Introduction, only the most relevant to this work will be detailed.

Among them, it is possible to divide open data criticism in two categories: (i) *Problems*, i.e., technological and political implementation failures that prevent open data to achieve their desired goals, and (ii) *Perils*, i.e., unexpected outcomes of open data implementation that are negative for the society as a whole, or specific groups.

Regarding the Problems, [Zuiderwijk et al. \(2012\)](#), in a very complete work, 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 open data description and searchability are great challenges in the field.

Regarding metadata structure, the Data Catalog Vocabulary (DCAT)⁶ ([CYGANIAK; MAALI; PERISTERAS, 2010](#)) was developed to provide standardized metadata for open data catalogues. If DCAT succeeds in providing a standard structure for describing open datasets, e.g., defining fields such as `dcat:title`, `dcat:description`, `dcat:keyword` and `dcat:theme`, harmonisation of the content of these fields is out its scope. This means that, if we have for Dataset 1 `dcat:theme spending plan`, and for Dataset 2 `dcat:theme budget`, both are not linked, although they are dealing with the same subject.

This issue can be tackled by enhancing metadata content, which is currently heavily influenced by the Linked Open Data (LOD) paradigm ([BERNERS-LEE, 2006](#)). In short, this approach aims to semantically enrich data by giving unique identifiers (Uniform Resource Identifier - URIs) to elements of a dataset, and linking these identifiers to commonly agreed knowledge bases, or web ontologies.

According to several authors ([SPECIA et al., 2007](#); [LIMPENS; GANDON; BUFFA, 2013](#); [ANGELETOU, 2008](#); [Van Hooland et al., 2013](#)), the procedure for enhancing metadata can be roughly divided into three stages: (i) *Metadata Cleaning-up*, i.e., spell-checking, equalising case and special characters, normalising gender and plural variations, and others; (ii) *Metadata Reconciliation*, i.e., matching metadata values with standard vocabularies, thesaurus or ontologies; and (iii) *Metadata Enrichment*, i.e., discovering meaningful relationships between several metadata.

Besides the data description challenge, another commonly cited impediment to the use of open data is the individuals and groups’ lack of capacity for dealing with open data. There has been a recently growing consensus on defining the skills of consuming, publishing and understanding data under the concept of *Data Literacy*.

As observed by [Bhargava and Ignazio \(2015\)](#), one of the first mentions to 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

⁶ Available at <http://www.w3.org/TR/vocab-dcat/>

reinforced the importance of teaching these three literacies for “students who need to critically evaluate information in arguments” (SCHIELD, 2004, p.1).

Although not directly mentioning the term Data Literacy, the previously cited collection of open data impediments (ZUIDERWIJK et al., 2012) dedicates a section for problems related to *understand ability*. Among them we find, for example, “Lack of skills and capabilities to use the data” and “Lack of knowledge about how to interpret the data”, which relate directly to the topic of Data Literacy.

1.3 Hypothesis

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

H1: *Cleaning up, reconciling and enriching metadata leads to a higher searchability of open datasets.*

The hypothesis puts in evidence the open data description problem and its relation with making it more easily available to the general public. It assumes that, if by some means, description of open datasets could be enhanced by cleaning up, reconciling and enriching metadata, open datasets would be more easily searchable by a broader audience.

The first part of H1 – *Cleaning up, reconciling and enriching metadata* – assumes that open datasets stored in ODPs exist, and that they are not adequately described by their associated metadata. From the first section of this Introduction, the existence of a considerable amount of ODPs is clear. Description problems were already cited above, and will be clarified during this text, specifically in [Section 2.7](#) and [Chapter 4](#).

The second part of H1 – *a higher searchability of open datasets* – is related to people wanting to search (and find) open datasets. The first assumption is that data consumers exist, i.e., that there is a demand from the society for open data. Several works point out the missing focus on users, or data consumers. Topics regarding open data demand and motivations are detailed in the next chapter, but it is worth mentioning here the work by [Davies \(2012\)](#). It describes a survey about the different uses of open data, and gives a more realistic perspective on how people really use data. The second assumption is that people currently have problems in searching for open datasets. Reports of problems on finding open data are also available on the literature. [Zuiderwijk et al. \(2012\)](#) summarizes 6 categories of find ability impediments, collected from the literature, workshops and interviews. This will also appear as a result of [Chapter 3](#).

Hypothesis H1 goes only until the point when data is found. But what happens after users get access to the desired open dataset?

Enhancing the description of open datasets will not make them more accessible alone. Increasing the level of data literacy on the society is also of paramount importance in order to democratise access to open data, and their claimed benefits. Otherwise, there is a serious risk of creating a “Data Divide”, i.e., extending the inequalities of our society to the open data access and thus *empowering the empowered*, as stated by [Gurstein \(2011\)](#). Although not formulated as a hypothesis, mainly because of our lack of instruments to validate it, we assume in this thesis that an increase on the collective level of data literacy could lead to a democratisation of access and benefits of open data. Thus, as a motivation of our work, we understand and assume the importance of gradually moving from a current scenario where the society is exposed to data filtered and explained by intermediaries, to another where people can critically read, understand and even produce data themselves.

1.4 Objectives

Our aim in this thesis is to develop an approach to enhance the description of open datasets, with the perspective of facilitating access to open data, and consequently improving the realisation of its benefits in a democratic way. In order to accomplish this main target, we formulate as specific objectives (SO):

1. To provide an overview on the state of the art regarding the open data landscape;
2. To investigate in which extent Data Literacy efforts can contribute to democratisation of access to open data;
3. To systematise literature efforts on cleaning up, enriching and reconciling open data descriptors;
4. To analyse the state of the practice on metadata usage on the ODP context;
5. To develop an approach for helping ODP managers cleaning up, enriching and reconciling their data descriptors; and
6. To develop an approach for semantically connecting ODPs, enabling an integrated and meaningful data search.

In the following section, the approach for reaching those targets is explained.

1.5 Solution Approach

In order to test the proposed hypothesis, and to accomplish our main objective, we propose the Semantic Tags for Open Data (STODaP) approach. As detailed in [Chapter 5](#), the STODaP approach consists in merging local strategies (at the ODP level) for cleaning up metadata and enhancing their quality, and global strategies, for reconciling metadata and providing a common environment for accessing open datasets.

Open Data Portals are collections of datasets, which have a set of metadata to

describe it. A widely used type of metadata are tags, free-text labels that can describe topics, geographical regions, temporal informations or other aspects regarding datasets.

Our approach consists in cleaning up, semantically lifting and reconciling these tags so that datasets from different portals, eventually in different languages, can be connected if their content refers to the same subject. By using the STODaP approach, users can find data about specific topics from several sources in one single place.

In order to specify and develop this main objective, SOs need to be accomplished. SO1 to SO4 are aimed at understanding our problem from its several perspectives, respectively the general open data landscape, the voice of (potential) open data users, related works about open data description, and the actual use of metadata in ODPs. SO5 and SO6 are subdivisions of the main product of this thesis - the STODaP approach.

1.6 Methodology

The personal motivation for this work comes from my previous experience in developing information systems for social movement activists, such as mapping systems, social networks and data analysis systems. This practical experience showed that a recurrent error is to look only to strict technical problems, ignoring influences of the social context. A learned lesson is that socio-technical problems need socio-technical approaches in order to be effective.

After attending to the Ph.D. seminars, the question of open data raised my interest, as being a technical challenge with a great potential of impact on the society. This process led to the general definition of the topic, and the specific problems to be dealt with.

The first methodological step after broadly defining the problem was a literature revision on open data. With this knowledge in hand, it was necessary to go to the field in order to experiment open data in practice and to hear potential users. Thus, theory about Popular Education and Participatory Research methods were reviewed, and a data literacy course was developed and presented, as well as the results of an associated participatory research.

As a result of this process, the problem definition and the solution approach were refined in order to fit the findings of the participatory research. Thus, it was also necessary to refine the literature revision, and also to look at the reality of metadata usage in Open Data Portals.

With all these components in hand: (i) a broader literature review; (ii) a field research; (iii) a narrower and deeper literature review; and (iv) a reality analysis, it was possible to dive into the development of the solution and conduct its evaluation. Finally, the whole process was discussed, and we analysed in which extent our results could be

generalised in face of the driven experiments, and if our hypothesis was fully validated.

1.7 Structure of the thesis

The remainder of this thesis is organised 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 organisation and linking possibilities on Open Government Data Portals. Some parts of this chapter are based upon the work published in [Tygel et al. \(2016a\)](#).

[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 organising data in ODPs appears as an outstanding impediment for using open data. In this chapter, we will use the results published in [Tygel, Campos and Alvear \(2015\)](#) and [Tygel and Kirsch \(2015\)](#).

[Chapter 4](#) sets the theoretical and practical basis to our solution. It 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, especially when dealing with tags. We also drive an analysis of metadata usage in open data portals.

[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 organisation 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. In this chapter, we will benefit from the work published in [Tygel et al. \(2016b\)](#).

In [Chapter 6](#) we drive an evaluation of the proposed approach. A theoretical background about search engine evaluation is reviewed, and we present our methodology. The experiment involving 34 participants is presented, highlighting results which show that STODaP helped participants to complete tasks quicker and preciser than using other methods.

[Chapter 7](#) presents the concluding remarks of this thesis, highlighting our contributions, stating the limitations of our approach, and 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 regarding transparency and citizen participation. Open data is even 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 the movement, highlighting its problems and selecting the main 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 web platforms, practitioners reports and official documents, seeking to reflect more clearly the current open data scenario. Rather than exhausting the topic, the aim of this chapter is to justify the importance of open data, present the open research issues and indicate the solution paths to be presented in the following chapters.

The chapter starts with a section dedicated to the alleged motivations for opening data, not only in the government context, but also within the research field. 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 [Section 2.5](#) we selected one special type of data – budget data – to describe in a more detailed fashion. The chapter continues with [Section 2.6](#), that seeks to analyse open data efforts in terms of impacts evaluation and value creation. Of crucial importance is [Section 2.7](#), where the problems of open data are analysed. This section is followed by a presentation of Linked Open Data approach ([Section 2.8](#)), regarded as a way to overcome some of the mentioned impediments. We finally conclude this chapter by pointing to selected references for a deeper understanding of open data.

2.1 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 organisation 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 and 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 the only one 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;
- 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.

Though less common, private companies may also be motivated to release open data for a number of reasons. For The World Bank, private companies in some sectors may release data to increase levels of customer engagement and loyalty (The World Bank, 2014). According to this report, companies can even release open data and charge for premium services and consulting. Benjamin Herzberg, in his blog also at The World Bank portal¹, lists a series of benefits for companies publishing open data: more efficient internal governance frameworks, enhanced feedback from workers and employees, improved traceability of supply chains, accountability to end consumers, and better service and product delivery. He argues that open data for private sector impacts both the bottom line and generates governance, environmental and social gains.

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 that the roots of open data praxis come from the scientific community, who first realized the importance of opening and sharing data. He argues 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 free². However, with the popularization of the Web, the increased speed in transmission rates, and the widely spread concept of Web Applications, it was recognized that opening the source code was not enough to guarantee 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

¹ Available at <<http://blogs.worldbank.org/voices/next-frontier-open-data-open-private-sector>>.

² 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/>>.

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 later, in 2011 the Open Government Partnership (OGP) was launched by eight countries, aiming to be a platform for national governments willing to be more open, accountable, and responsive³. 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 Charter⁴ by the G8 leaders, in 2013. The charter is based on six principles to be followed the 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 Data⁵, 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.

³ Available at <http://www.opengovpartnership.org/>

⁴ Available here: <http://opendatacharter.net/>

⁵ Available at <https://opengovdata.org/>

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 usability 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 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. Germany⁶ and UK⁷. The Open Data Commons platform⁸ 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). Therefore, it has to be clear that “License-free” means *possessing a license which guarantees freedom of use*, and not *free of license*.

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, i.e., data must be subject to public contestation; and
14. Interagency coordination: This means “developing a shared data standard, or adopting an existing standard, possibly through coordination within government across agencies”.

⁶ Available at <<https://www.govdata.de/dl-de/by-1-0>>

⁷ Available at <<http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>>

⁸ Available at <<http://opendatacommons.org>>.

Another widely accepted definition comes from the general Open Definition, which is currently on version 2.1⁹. In contrast to the previous definition, this one is not aware of 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).

linked open data; a step further to open data which exploits the relations between different data.

In order to help publishers in creating a roadmap towards the implementation of the Linked Open Data (LOD) paradigm, [Berners-Lee \(2010\)](#) proposed a five stars schema, where each star represents one step further in turning data more open and linked. The scheme starts from a PDF file with open licence, symbolizing a reusable closed date, and finishes with LOD, the datasets have unique identifiers and are linked to other datasets through the LOD cloud. LOD will be further detailed in [Section 2.8](#). The key elements for each of the five stars are:

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

Though not cited so much, the Three Laws of Open Government Data developed by [Eaves \(2009\)](#) have 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; and
3. If a legal framework doesn't allow it to be repurposed, it doesn't empower.

Finally, some recent works defined intermediate levels between closed and open data. The Open Data Institute defines a Data Spectrum, which ranges from closed, through shared until open data. [Figure 1](#) pictures this definition.

With the same idea of defining intermediate levels of data openness, [Bargh, Choenni and Meijer \(2016\)](#) proposed the Semi Open-Data Paradigm. The objective is the analyse

⁹ Available at <http://opendefinition.org/od/2.1/en/>

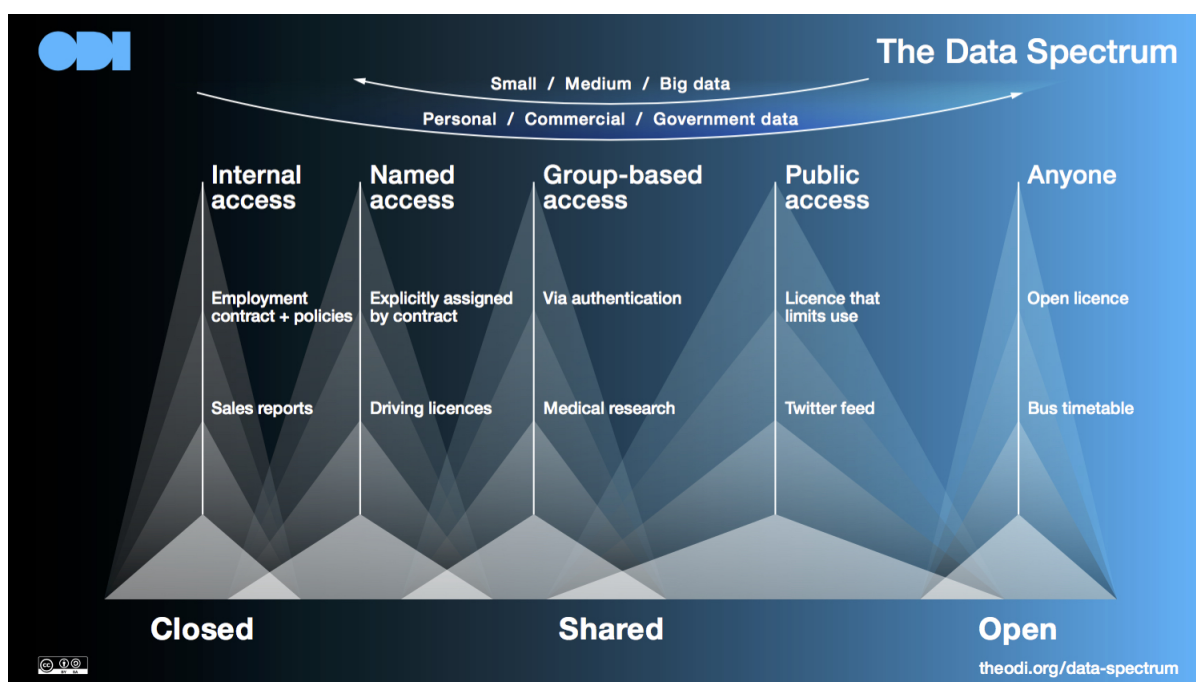


Figure 1 – Data Spectrum as a definition of steps between open and closed data. Source: <http://theodi.org/data-spectrum>

the dissemination of data through several dimensions, as publicity, completeness, timeliness, metadata, and other. For each dimension, several levels 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 organisation’, ‘share data within an organization /ministry’, and ‘share data among a federation of organisations’ and finally ‘share with the public’.

2.4 Open Data Landscape

While the number of open data initiatives around the world increases dramatically every year, several research projects driven from academia and/or civil society organisations 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: Open Data Index is one of the most important platforms for analysing the worldwide open data landscape. Evaluations started in 2013, when the Open Data Index analysed 60 countries. In 2014 this number grew to 97, and in 2015 it covered 122 countries. 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: Open Data Barometer also focuses 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, beyond 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, data about public transportation positioning, 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 criteria¹⁰, 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 (VAFOPOULOS 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 (VAFOPOULOS et al., 2013). The absence of standards can lead to data misuse (ZUIDERWIJK; JANSSEN, 2014b), 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 the framework as a tool to design more adequate budget publishing systems. Together with other ongoing initiatives (OPENSPELLING, 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.

¹⁰ Other criteria can be found at <<http://bit.ly/1929F11>>.

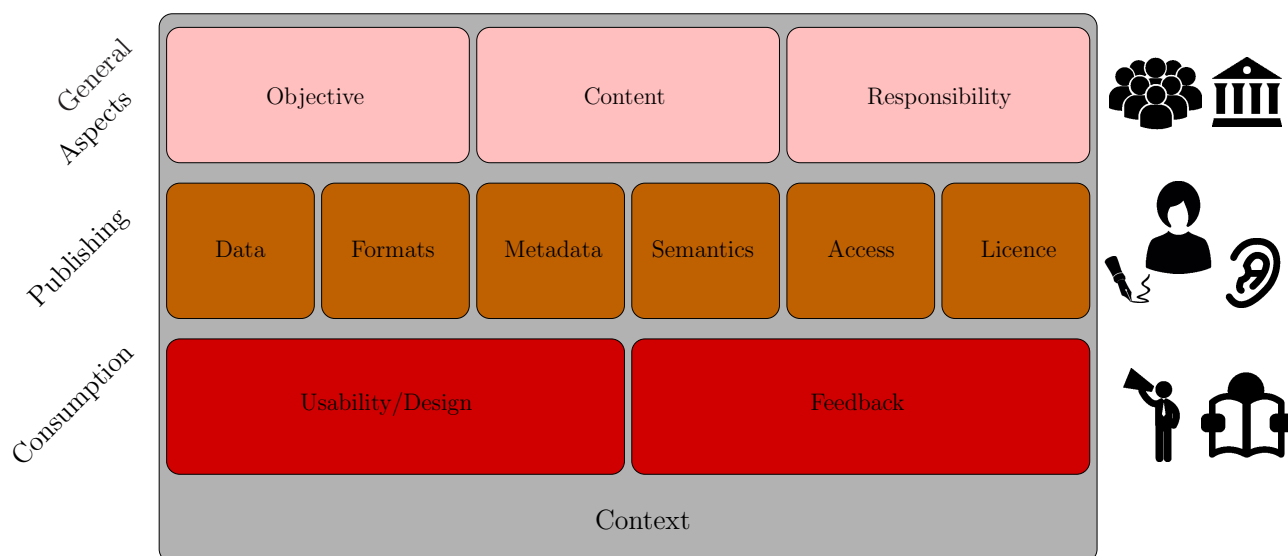


Figure 2 – Analysis framework for open budget initiatives. The four parts – General Aspects, Publishing, Consumption and Context – are interconnected, and composed by several dimensions. Icons:Flaticon (CC).

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*¹¹.

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 organisations. 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.

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.

¹¹ Available at <http://community.openspending.org/map-of-spending-projects/>.

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

2.6 Evaluating Open Data Impacts and Value

As seen in [Section 2.4](#), mapping open data initiatives is a very important way of assessing the amount of published data between countries. However, 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 and 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.

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 and 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 and 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.

Regarding open data economic value, some reports sought to estimate the amount of money that could be unlocked by open datasets. According to [Manyika et al. \(2013\)](#), this value would be between U\$3 trillion to U\$5 trillion annually, mainly on using open data to analyse consumer preferences and allow companies to improve their offers. A compilation of several estimations by The World Bank ([The World Bank, 2014](#)) presents most values around tens to a few hundreds of billions of dollars. According to them, there are certain types of data which possesses a higher potential for generating economic value,

such as geospatial reference data, weather, road and transports, and company registers. The same report highlights four examples of companies that recently grew upon open data, on topics home and real estate (Zillow and Zoopla), land registry, maps (Waze) and soil quality based on weather observations (Climate Corp). Still according to the World Bank, there are several categories of open data based businesses, such as: suppliers (that charge only for special services, not for data), aggregators, developers, enrichers and enablers.

2.7 Problems of Open Data

Although the vast majority of research about open data assumes that publishing public data in open formats will bring mostly good impacts, a number of recent works are dedicated to show the other side. In this sense, a good starting point are two works from the same research group that make an in depth research on the problems and negative effects of open data.

In the first one, [Zuiderwijk et al. \(2012\)](#) analysed the socio-technical impediments that hinder 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 and Janssen \(2014b\)](#) 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 and Schossböck \(2016\)](#) interviewed public servants in German speaking regions in order to gather barriers for the implementation of open data in the public sector. As result, three main impediment classes were found: information

cultures and divergent interests in agencies, limited innovation potential in organisational cultures and limited communication of strategies. Authors conclude that “if organizational aspects of information culture are not addressed, the value of Open Government might not be understood” and that “cultural and organizational factors, as opposed to a simple lack of knowledge, play a crucial role regarding the implementation of open technologies in the German-speaking region.” (PARYCEK; SCHÖLLHAMMER; SCHOSSBÖCK, 2016, p.7).

Another aspect revealed by [Zuiderwijk and Janssen \(2014b\)](#) is the lack of informations about the actual use of open data by people: “The interviewees stated that they do not have much more information about how the open data have been used other 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 and non-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; and
- 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, the author cites some challenges in the social and technical fields. The priority, according to the author, 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 information 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

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/odl/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 \(2010\)](#)

discussion and interaction around them.

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.
- Expert advice is needed to use the data.

Zuiderwijk and Janssen (2014b), 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, 2014b, p.150).

Inequality 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, p.2).

A data divide between countries is also mentioned as one of the conclusion of the Open Data Barometer project. Davies, Sharif and 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.”

In order to be usable, published open datasets have to be described in a logical way that allows people to find it. Regarding open data repositories, the description challenge has an internal aspect, i.e., the way each ODP describes their own data, but also a global aspect, regarding to a harmonization among different repositories. Both issue are directly related to the proper use of metadata.

The impediments compilation by Zuiderwijk et al. (2012) cites some related problems, such as “absence of commonly agreed metadata”, “insufficiency of metadata”, “the lack of interoperability” and “difficulty in searching and browsing data”. The same study

also states that “A lack of open data standards between (levels of) government organizations has been identified as a barrier to open data usage by citizens and businesses and subsequently new open data policy”.

When analysing open data of different contexts, language aspects quickly emerge: “Language barriers and interoperability aspects need to be tackled so that information resources from different organizations and countries can be combined” (ZUIDERWIJK et al., 2012).

Description of datasets is also problem for agents in the private sector willing to use open government data. According to Roseira (2016), firm managers interviewed point out that most datasets have incomplete or non-existent metadata. This lack of data quality generates a higher workload on cleaning and harmonizing data. The study included that firm managers desire to see advances on datasets standardization in order to boost open data economic value creation at national and international levels.

2.8 Linked Data towards Semantic Description of Open Data

One of the strategies for adding value to data is interlinking it with other datasets. As described by Attard, Orlandi and 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 description 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 forming a graph structure. 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 platform¹², 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 and 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, of 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 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.

¹² Available at [<http://lod-cloud.net/>](http://lod-cloud.net/).

2.9 Conclusions

In this chapter, an overview about Open Data was presented. A historical view was emphasized in order to better contextualize this movement. We highlighted some aspects as definitions and main research problems currently posed. Linked Open Data was presented as a possible way to overcome some of the problems, while turning data more accessible and meaningful.

It is clear from this overview that open data is definitively inserted in the public policy agendas. Pressures come both from the top, pushed by multilateral organisms such as G7, European Union and Open Government Partnership, and from the roots of civil society claiming for transparent and accountable regimes.

While research regarding problems on the implementation of open data are important, early detection of perils of this trend is of utmost importance in order to avoid or minimize open data misuse. In this sense, the risk of enlarging social inequalities and creating a data divide must be seriously considered by public agents in charge of implementing open data policies.

For a more complete view on Open Data, please consult the following selected bibliography:

- *Community Informatics and Open Government Data*, by [Davies and 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 mostly after literature research. In the next chapter, a field research is driven via open data classes for social movement activists, where impression could be taken from the real users.

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 organisations. It is fundamental to guarantee equal opportunities for learning data skills in order to avoid enlarging the data divide, as mentioned in [Section 2.7](#).

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 driven as 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 (HUIJBOOM; BROEK, 2011). Meanwhile, the severe social inequalities faced all over the world, reflected directly in the education level of the population, are creating 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 First Data Literacy Workshop, co-located at the 2015 ACM Web Science conference. In one of the published papers, Bhargava and 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.

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).

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

Wolff, Kortuem and Cavero (2015) describe 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 and Patton (2006) enforce the difference between statistical and data literacies: while the first one concentrates on applying statistical methods to data, the second one is more concerned 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 organisations, journalists and citizens with the skills they need to use data effectively”, under the slogan “Evidence is Power”. In 2014, the School of Data organised 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 organisation, 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 and 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 and 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 and Kirsch (2015)

president of Brazil, João Goulart, invited Paulo Freire to organise 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, when Paulo Freire was developing his method, the official 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 the official 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

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.

process must start from the educands reality. Thus, there is a commitment for educators 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, organises 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 organisation 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 was latter included as a fourth stage in the literacy method:

3.2.1.4 Systematisation Stage

In this moment, the lived experience are organised, 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

We here discuss the parallels between both literacies, and the possible contributions of Paulo Freire to the topic. Finally, we derive our own definition of Data Literacy in the end of [Subsection 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
Problematization	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	Organisation, interpreting, and presentation of the lived experience	Organising 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. When immersed 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 built 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 organisation, 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 moment 1, 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 the moment 2, data from this context is gathered. The critical analysis of this data (moment 3) is necessary in order to understand which

perspectives are illuminated by this data, and which are hidden. Finally, at the moment 4, after the critical analysis of data, it is possible to look again to the context, 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.

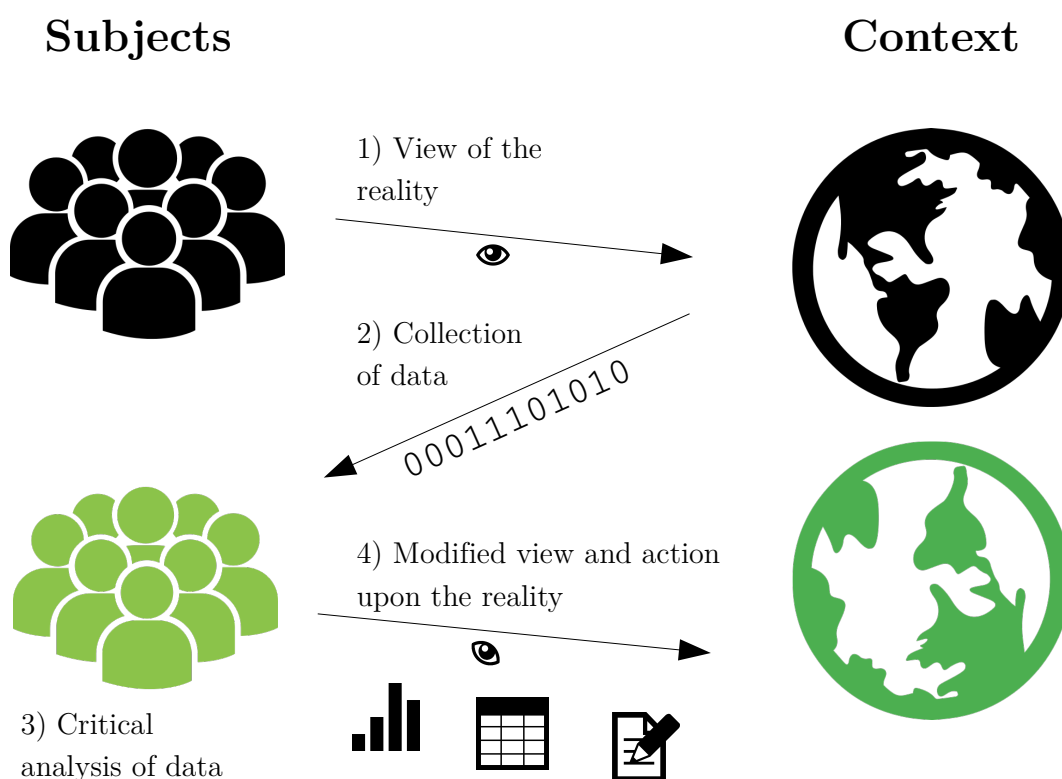


Figure 4 – Critical data literacy process. Source: [Tygel and Kirsch \(2015\)](#)

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 and 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 alluring 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 organisations? What was the political position of that organisation 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 organised, 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:

Definition: *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 organisations, 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 section, 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 include deriving more tangible examples of the application 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 democratisation 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

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

computer (which could also be offered by the organisers). Good quality internet access provided by the organisers is also highly desirable.

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 Table 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.

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)

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 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, students 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.

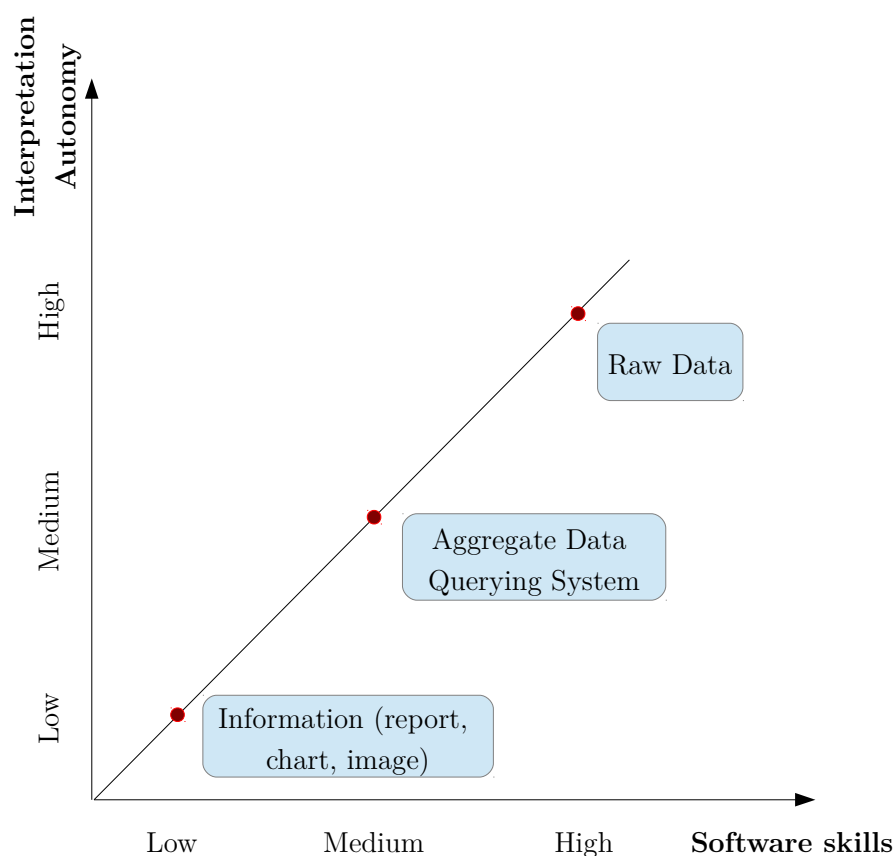


Figure 5 – Trade-off between interpretation autonomy and software skills needed.

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), 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

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

Initiative	Publisher	URL
Environmental Conflicts Maps (Brazil)	Fiocruz and FASE	< http://www.conflitoambiental.iciet.fiocruz.br/ >
Environmental Conflicts Maps (Minas Gerais, BR)	Federal University of Minas Gerais, Brazil (in partnership with a number of social movements)	< http://conflitosambientaismg.lcc.ufmg.br >
Atlas of Environmental Justice	23 worldwide organisations	< http://www.ejolt.org/maps/ >
Agroecology initiatives	ANA/ABA (Brazilian National Social Movements related to agroecology)	< http://agroecologiaemrede.org.br >
Land Conflicts	Comissão Pastoral da Terra (CPT)	< http://cptnacional.org.br >

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 Electoral Justice, a civil society organisation 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 organised 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 Conflicts. 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 Conflicts 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, organised 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 organisers. 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 and 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 [Apêndice B.](#), in Tables 22, 23, and 24.

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 and 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 conflict. 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 organise 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 22, 23, and 24 (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

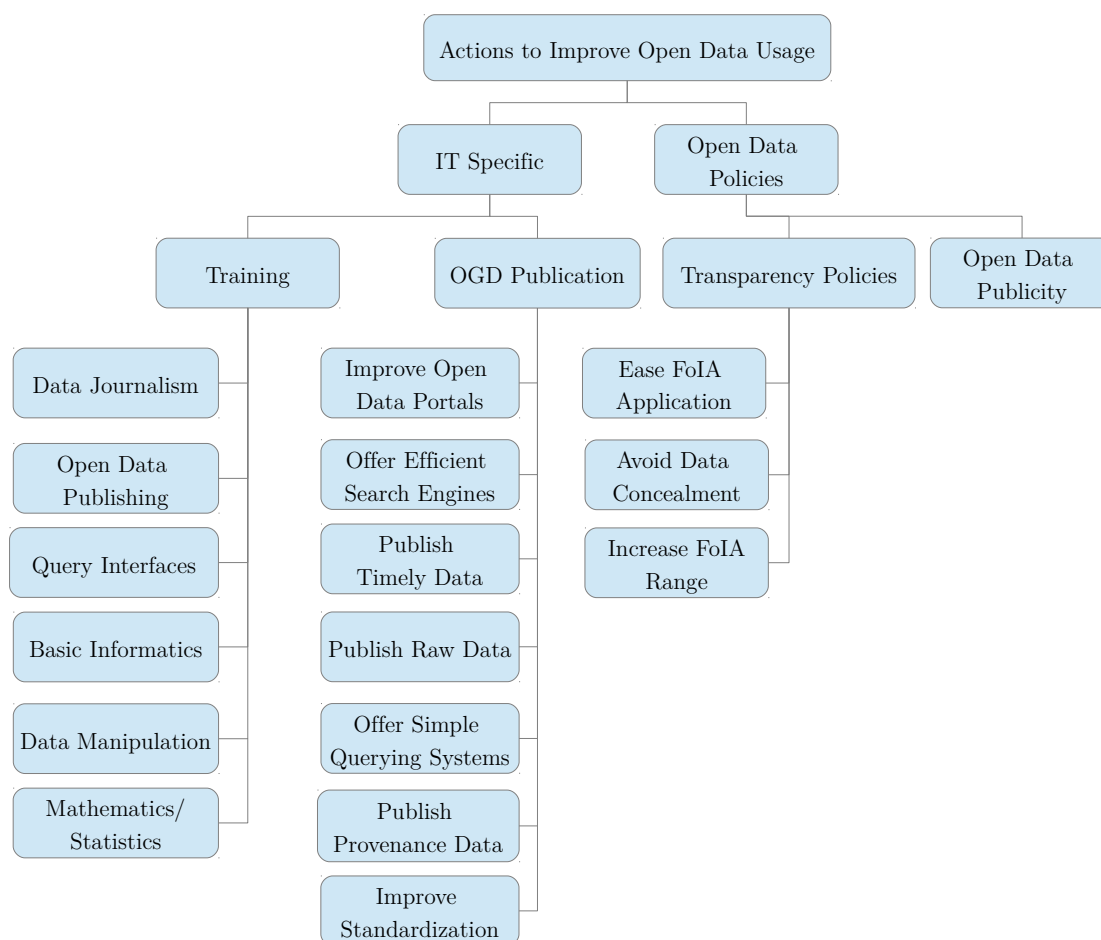


Figure 6 – Classification tree for open data engagement actions, systematized from Tables 6, 7 and 8. The first classification is a distinction between Information Technologies (IT) Specific and Open Data Policies related issues. There is no intention to imply a duality between social and technical issues, 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 22, 23, and

24).

3.5 Conclusions

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 description is an important question that hinders the achievement of open data promises, the next chapters we will be dedicated specifically to this topic. The following chapter presents a literature review on methods for dealing with semantic descriptors for open data, and an analysis of the use of metadata in ODPs. The intention is to prepare the ground for presenting the Semantic Tags for Open Data – STODaP – approach, in [Chapter 5](#).

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 pointed out a significant problem related to the description of ODPs. Following this motivation, we investigate in this chapter the issue of metadata for open data. The chapter is divided in two parts, covering:

- A literature investigation over the possible strategies to deal with semantic metadata for description of open datasets;
- An analysis of the current usage status of metadata in ODPs.

Particularly, in the first part we select works related to semantic enrichment of metadata in ODPs, in order to position the main contribution of this thesis presented in the following chapter. This 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 topic, highlighting the main related works, their gaps and relations to this work. We start with Assessment of Metadata in [Subsection 4.1.3](#), followed by Metadata Cleanup in [Subsection 4.1.4](#), Metadata Reconciliation in [Subsection 4.1.5](#) and finally with Structure Emergence in [Subsection 4.1.6](#).

The second section aims to bring light over the current status of metadata usage in Open Data Portals. Based on the CKAN Census of ODPs, we profile 87 portals and analyse several aspects regarding metadata. The analysis embraces not only local aspects regarding individual portals, 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](#)).

We conclude in the last section with an evaluation of the literature gaps and the actual problems detected in ODPs, pointing out our strategies to be developed in the next chapter.

4.1 Semantic Metadata: A Literature Review

4.1.1 Introduction

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 such as clean, well organised,

detailed, complete, accessible, and meaningful. Intuitively, metadata are meaningful if they bring new information – meaning – for data. If a consumer asks: “Which bananas 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 of origin, among others.

In 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 2.8](#). This procedure is also called Semantic Enrichment or Semantic Lifting. A special type of metadata is lately of particular interest for semantic lifting: tags, free-text labels that can describe several aspects of data. There are two particularities regarding tags that make them interesting for semantic processing: (i) the ease of input, because there are usually no constraints for users typing tags; and (ii) the social architecture, that allows different users to tag the same data element. Combination of both aspects results in large social tagging sets, which are also called folksonomies. [Limpens, Gandon and 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 hyponyms¹;
2. suggesting related tags to extend the search;
3. semantically organising tags to guide novice users in a given domain more efficiently than with flat lists of tags or occurrence-based tag clouds; and
4. assisting disambiguation.

A more detailed view about problems caused by the absence of semantics in metadata is described by [Marchetti and Rosella \(2007\)](#). According to the authors, 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’);

¹ In linguistics, hyponyms are words that share the same type-of relationship with an hypernym. Using the bananas example, Cavendish and Latacan are hyponyms because both are types of bananas, their hypernym.

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, p.2).

Around ten years ago, discussions 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 of 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 contradictory: 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, by allowing the community to give a negative polarity for a tagging made by a spammer.

Another important work introducing this topic is "*Ontologies are us: A unified model of social networks and semantics*" (MIKA, 2005), also published first in November of 2005. Mika also disagrees that ontologies and folksonomies are contradictory, but differently from Grubber, for who both are distinct concepts (Apples and Oranges) 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 and Yu (2006) added a time dimension to the tagging model. And to the best of our knowledge, Newman (2005) was the first to propose an ontology for tagging. This work was further extended by Knerr (2006), who proposed the Tagging Ontology, depicted in Figure 7. All dimensions proposed by Wu, Zhang and Yu (2006) and Grubber (2007) are present and further detailed in this ontology. The central element is the tagging, which acts as an event joining a tag, a tagger, a domain and a resource, as well as other optional attributes. One of these attributes is `hasType`, which is designed for qualifying the tagging as video, image, audio and others.

Although crucial, these models are not aimed at solving the problem of creating domain ontologies emerged from collaborative tagging. Halpin, Robu and 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 organising the knowledge contained in the tagging activity. The Meaning of a Tag (MOAT)

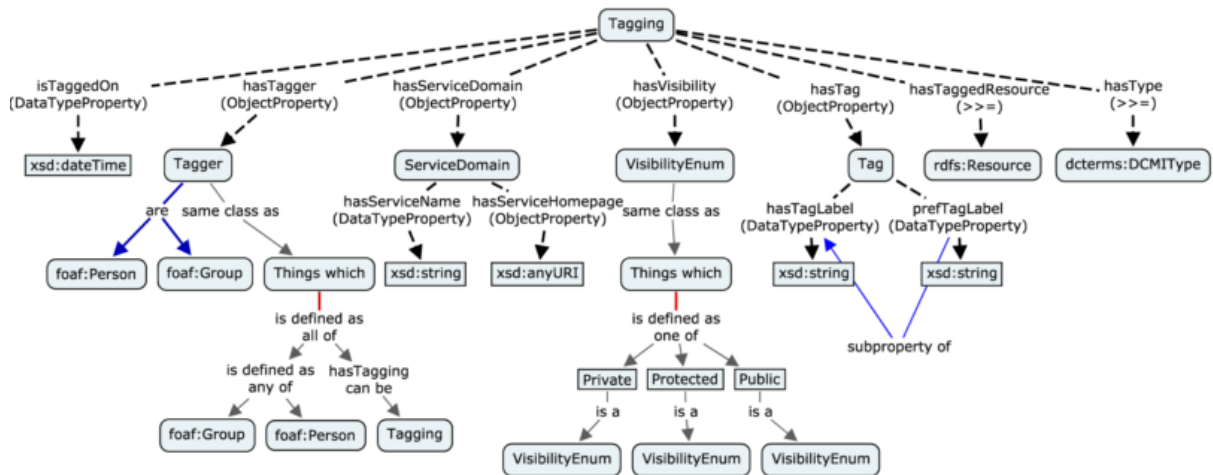


Figure 7 – Tagging Ontology. Source: Knerr (2006)

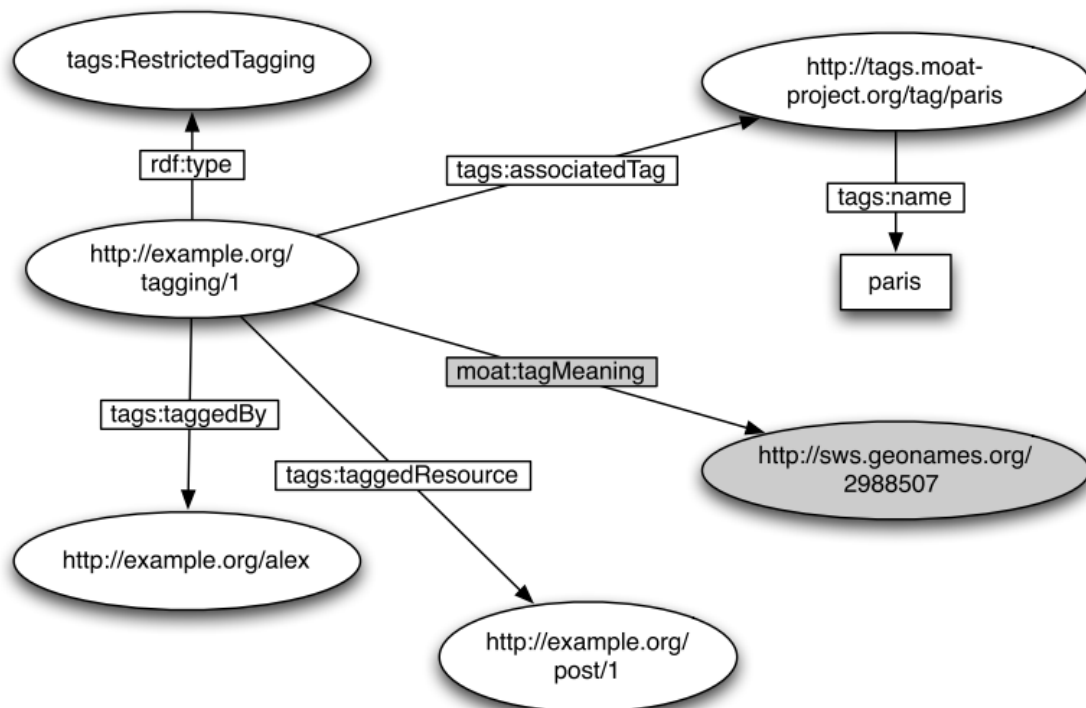


Figure 8 – Meaning of a Tag (MOAT) Ontology. Source: Passant (2008)

architecture was the first to explicitly include the concept of tag meaning, associating each tagging element to a LOD resource (PASSANT, 2008). Figure 8 shows an example of its application.

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). In the first work, seven models were compared, using as criteria their suitability

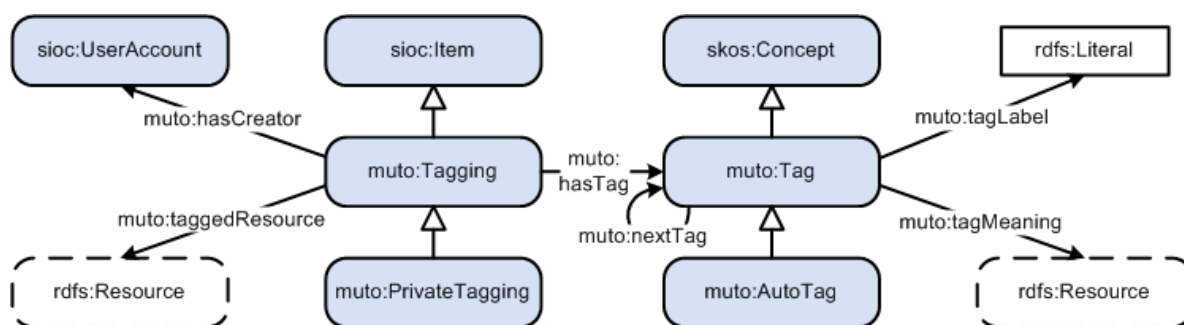


Figure 9 – MUTO Ontology. Source: Lohmann, Díaz and Aedo (2011)

to represent tagging activities, or to represent features of folksonomies. Authors conclude that SCOT² ontology is the only one that presents a higher level of sophistication in both directions. The most recent attempt to build a tag ontology is the Modular Unified Tagging Ontology (MUTO)³, shown in Figure 9, and described by Lohmann, Díaz and Aedo (2011). 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. An important highlight of MUTO is the possibility of connecting a meaning (an RDF resource) to a tag through the `muto:tagMeaning` property, as introduced by the MOAT ontology.

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 and Rosella (2007) point 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.”

4.1.2 Characterization of the Contribution

The main contribution of this thesis is an approach for cleaning up, semantically enriching and reconciling metadata descriptors of open datasets, with a special focus on tags. Thus, the following topics are considered to be related works, and will be analysed in the following:

- Metadata assessment: how to assess the use and quality of metadata;
- Metadata clean-up: how to enhance the quality of the metadata using strategies such as spell-checking, detection of similar words, special characters equalization, and others;

² Available at <<http://rdfs.org/scot/spec/>>.

³ Available at <<http://muto.socialtagging.org/core/v1.html>>

- Metadata reconciliation: how to align metadata with standard vocabularies, thesauri and ontologies;
- Structure emergence: how to find semantic relation between metadata.

Regarding the related bibliography, it is necessary to highlight that the vast majority of scientific works about tagging and semantics focus on a different context in relation to ours. It is normally assumed a folksonomy environment, where tags are attributed to resources by the crowd, passing through a crowd-selection mechanism, which may enhance the tagging quality, but can also insert some inherent noise. This is applicable to platforms such as del.icio.us⁴ or Flickr⁵, 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.

In the following subsections, we analyse the literature contributions regarding the above cited topics in relation to our work.

4.1.3 Metadata Assessment

An important step in working with metadata is to develop methods for evaluating quality aspects of it. Reiche and 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 and Duval (2006). Although the metrics definition are significant, their implementation in an automatic context is simplistic, which in practice do not make its use attractive.

In relation to the metrics for tagging environments, some related ideas could be found in the literature. For example, Umbrich, Neumaier and 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. This metric, however, is applied to file type metadata, and not 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 a word without being a hashtag and as a hashtag; consistency; and stability over time.

Colpaert et al. (2014) presented a method for calculating interoperability between

⁴ Available at <<http://del.icio.us/>>.

⁵ Available at <<http://flickr.com>>.

ODPs based on identifiers used in datasets. These identifiers are unique identifications for data items, and the process of considering them equal or different is manual. The metric verifies if the same identifiers were used to represent the same concepts in different datasets in order to calculate the interoperability metric.

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

4.1.4 Metadata Clean-Up

When dealing with metadata of large datasets, a cleaning up procedure is usually the first step before start working with them. There are several strategies for cleaning up tags described in the literature.

[Angeletou \(2008\)](#), in a context of semantic enrichment of folksonomy tagspaces, 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 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, their procedure to integrate folksonomies to the semantic web also includes a pre-processing phase. As in the previous work, the first step consists in removing tags with low chances of being mapped in an ontology. In the sequence, a series of heuristics are used to group morphologically very similar tags, including the Levenshtein distance ([NAVARRO, 2001](#)). In order to choose the most significant tag in a group, preference is given to terms 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 in an isolated form.

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 describe cleaning activities such as deduplication (remove duplicate entries), atomization (explode overloaded fields), applying facets and clustering.

As we can see, metadata clean-up mostly consists in finding representations that are more suitable to serve as input to the next processing step. The “cleaner” metadata is, the higher are chances of finding a commonly agreed meaning to it, as we will see in the following.

4.1.5 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 the Library of Congress Subject Headings (LCSH)⁶ and Powerhouse Museum Object Name Thesaurus⁷. 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 metadata curators to ensure the compliance of datasets with controlled vocabularies. Alongside the automatic procedures, users are allowed to build a synonym table in order to provide manual input to the algorithm.

A whole Semantic Tagging system is proposed by [Marchetti and Rosella \(2007\)](#). The system, implemented as a browser plugin, allows users to tag web resources and choose 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 and 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 hypernyms 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 Watson⁸ platform, and choosing the most adequate according to the defined criteria.

Instead of grouping tags using semantic criteria, [Specia et al. \(2007\)](#) use 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.

⁶ Available at <http://id.loc.gov>.

⁷ Available at <http://www.powerhousemuseum.com/collection/database/thesaurus.php>.

⁸ Available at <http://watson.kmi.open.ac.uk/WatsonWUI/>.

4.1.6 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 pairs of tags appear on the same ontology, and in case of success, relations are extracted directly from the ontology.

Several approaches described in the literature make use of similarity measures in order to determine the relation between two tags. The topic of similarity measures is very extensive, and several strategies can be found on the literature ([HARISPE et al., 2015](#); [HARISPE et al., 2014](#); [TRILLO et al., 2007](#); [CILIBRASI; VITANYI, 2007](#)). 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 words through the WordNet tree. It is worth highlighting a paper by [Cattuto et al. \(2008\)](#), where several measures of relatedness are compared to WordNet similarity 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 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.

A very interesting point-of-view on this topic is brought by [Limpens, Gandon and 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 10](#) 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.

4.1.7 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 DBpedia 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.

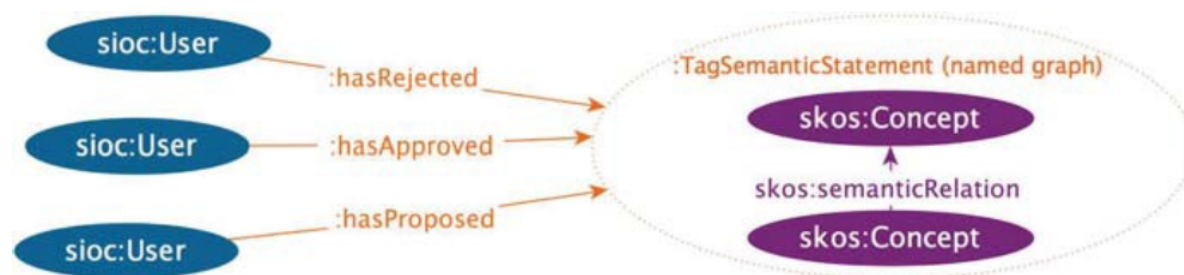


Figure 10 – SRTag RDF schema. Source: [Limpens, Gandon and Buffa \(2013\)](#)

[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.

4.1.8 Semantic Lifting in ODPs

The problem of semantic lifting in ODPs was tackled by [Ermilov, Auer and Stadler \(2013\)](#) and [Ding et al. \(2011\)](#). 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 CKAN⁹, the standard open-source software for ODPs. The CKAN community publishes a census¹⁰, 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

⁹ Available at <http://ckan.org>

¹⁰ Available at <http://ckan.org/instances>

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

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 online¹¹.

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 project¹².

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 organising ODPs – groups of datasets – Table 8 also shows the number of groups per portal, and the number of datasets inside each one. While the 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 organise data.

In the following, we present the metrics and their results divided into Local Metrics, i.e., applied separately to each portal, and Global Metric, where a joint analysis is driven. First ones aim to assess the use of tags and verify enhancement possibilities, and the last ones assess the viability of using tags as main elements of communication between portals.

¹¹ <<http://bit.ly/1NGygtk>>

¹² <<https://github.com/alantysel/StodAp>>

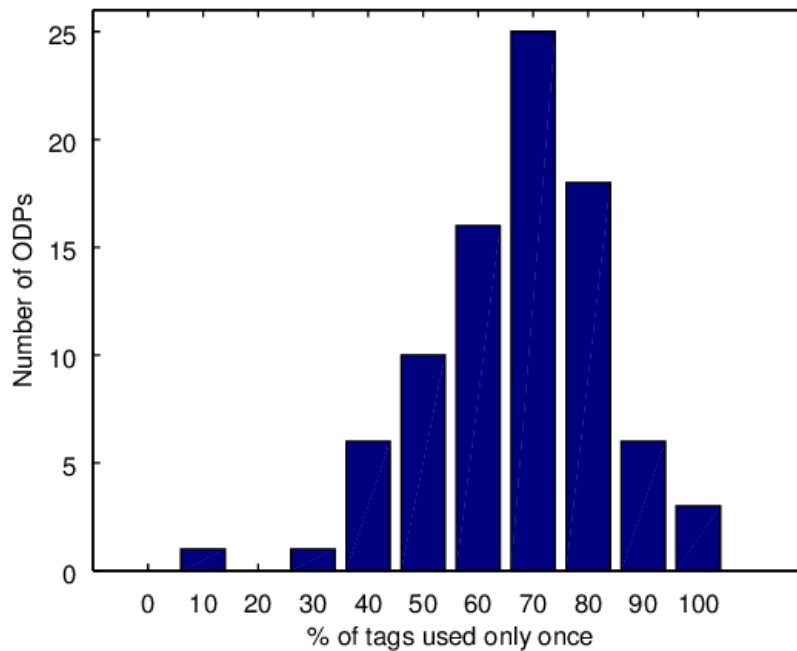


Figure 11 – Re-use of tags inside an ODP. The graphic shows the distribution of the percentage of tags used only once.

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 11](#) 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.

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 and Polleres \(2015\)](#), if the tag metadata is being actively used in the portals. We must note that the results of this metric cannot lead to further conclusions, since we do not intend to define an optimal value for the number of tags per dataset. Using few and consistently used tags may support the organisation of datasets better than many incoherently used ones. On the other hand, few tags may not label the content adequately.

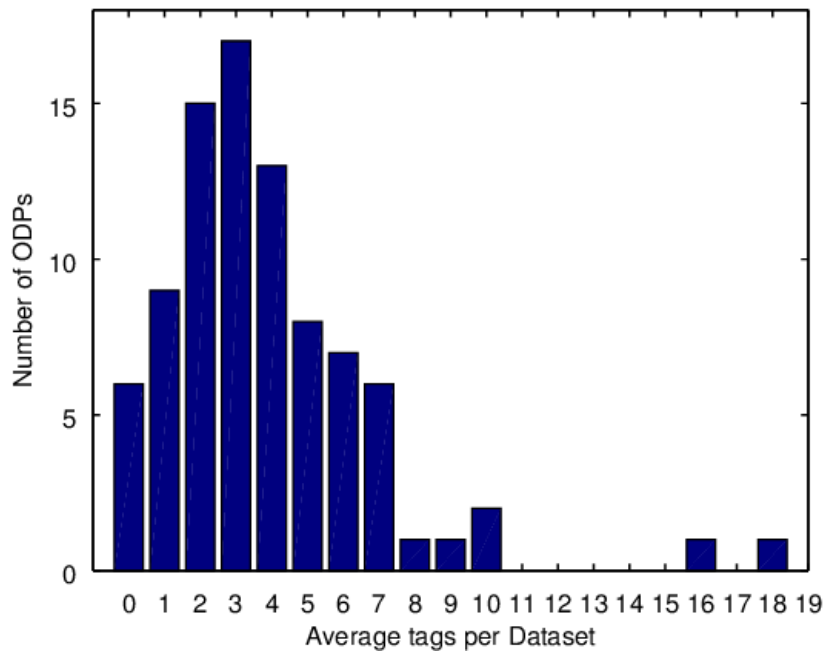


Figure 12 – Distribution of the average number of tags used per dataset in ODPs.

Figure 12 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`.

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.

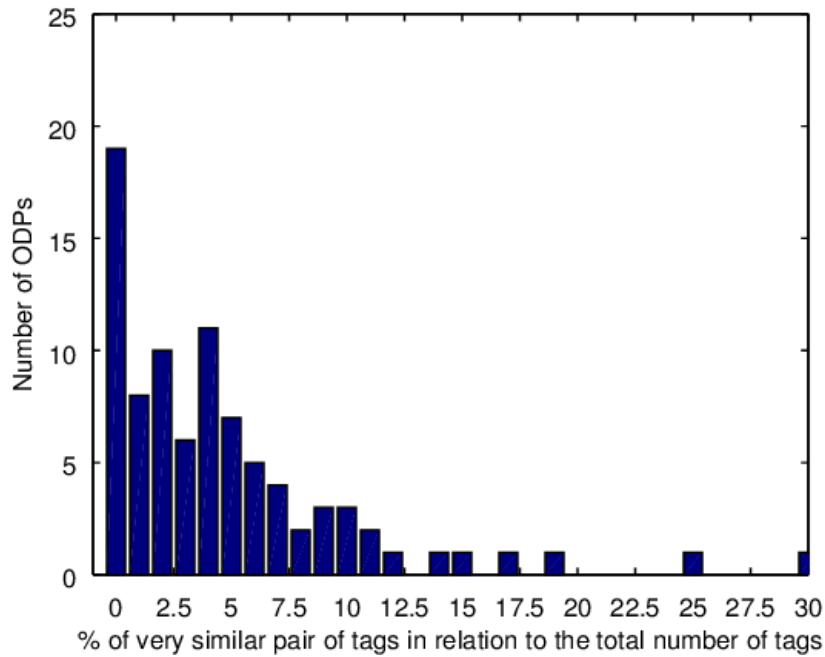


Figure 13 – Percentage of very similar tags in ODPs, where the difference lies only in capitalization or special characters.

For our purposes, we define similarity as:

$$S = \frac{\text{Similar Pairs}}{\text{Number of Tags}} * 100, \quad (4.1)$$

where Similar Pairs is the number of tag pairs where tags are equal after lowercasing and unaccenting. Figure 13 shows the distribution of similar tags inside each ODP. 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 87, revealed no similar tags at all. It should be noticed that these portals use far less tags (average 148 per portal) than the global average of 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 by [Umbrich, Neumaier and 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 [Chapter 5](#).

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. [Passant \(2008\)](#), while building the MOAT ontology¹³, 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 for data description – ontology – and for individual instances. DBpedia¹⁴ 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, 2015](#)), 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 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

¹³ <http://muto.socialtagging.org/mirror/moat.rdf>

¹⁴ <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 5.2.4, 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%

tags can return several meanings, such as `leaves`¹⁵, 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.

4.3 Conclusions

In this chapter, a theoretical and practical analysis about metadata and tagging in ODPs was driven. After a general overview on semantic tagging, a literature revision regarding metadata assessment, clean up, reconciliation and relationship presented the recent advanced of metadata curation in relation to the Semantic Web tendency.

Apart from the literature review, looking at the actual use of tags in ODPs was also necessary. An analysis of 87 ODPs revealed that: (i) tags in ODPs are widely used, but in a non-systematic way, which hinders the search ability of datasets inside it, 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.

Ideas and gaps noticed on the literature, and actual problems and potentials detected on ODPs are the basis for proposing the STODaP approach that will be described in the following chapter.

¹⁵ <<http://www.lexvo.org/page/term/eng/leaves>>

5 STODaP Approach

As observed in the previous chapter, literature related to semantic enhancement of ODPs metadata has still some significant challenges, such 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 those issues, we describe in this chapter the Semantic Tags for Open Data Portals (STODaP) approach for improving tag curation within and across ODPs, and linking ODPs via a common basis for semantic metadata¹.

Besides the comprehensive analysis of tag usage in 87 ODPs, shown in previous chapter, that justifies the need and benefits of better tools for managing tags, our main contributions with this approach are:

- An approach for cleaning and reconciliation of metadata in ODPs;
- An approach for semantic lifting of metadata in ODPs;
- A centralized repository for connecting ODPs through meaningful shared tags.

This chapter begins with a short motivation to the topic of our work. For a deeper analysis, readers are referred to the previous chapters. The main part of this chapter lies in [Section 5.2](#), where our approach for semantic tags in open data portals is explained. STODaP architecture is detailed, and every component and their connections are explained. Following that, [Section 5.3](#) presents the implementation architecture, detailing the technological choices used to implement the STODaP server and the associated plugins. We present afterwards some quantitative results in [Section 5.4](#), and finish with a conclusion regarding the developments presented in this chapter.

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) one question arises: how larger datasets can be published and described in order to make them easily discoverable and facilitate their integration as well as an analysis of their data?

¹ This chapter is an extension of [Tygel et al. \(2016b\)](#).

One approach for addressing the problem of data dispersion are data catalogues, which enable organisations 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 registry², for example, already lists 519 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 organisations and NGOs, as well as scientific data portals.

Based on a discourse of increasing transparency and citizen engagement, governments and public administrations all over the world are implementing ODPs. These ODPs comprise large amounts of data, mostly structured in the form of tabular data such as CSV files or Excel sheets. However, large quantities of PDF and other closed formats can still be found in ODPs. The aim of an ODP is to be a one-stop-shop for citizens and companies interested in using public data produced by government or a civil society organisation. Examples are the US data portal³, the UK data portal⁴, the European Commission⁵ 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 data analysis 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 portal is the Global Biodiversity Information Facility Data Portal⁶. Also many international and non-governmental organisations operate ODPs such as the World Bank Data Portal⁷ or the data portal of the World Health Organisation⁸. Although being a relatively new type of information system both commercial (e.g. Socrata) and open-source (e.g. CKAN) data portal implementations are already available.

In an ODP, metadata used to describe datasets comprise normally title, description, last update, format, size, license, and categories or groups, but most importantly free-text words or sets of words used as labels – the so called 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

² Available at <<http://datacatalogs.org>>.

³ Available at <<http://data.gov>>.

⁴ Available at <<http://data.gov.uk>>.

⁵ Available at <<http://data.europa.eu/euodp/>>.

⁶ Available at <<http://www.gbif.org/>>.

⁷ Available at <<http://data.worldbank.org>>.

⁸ Available at <<http://apps.who.int/gho/data/>>.

folksonomies – because all users are 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 administrators 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. The concept of allowing web objects to be easily tagged and retrieved using these tags, however, remains the same.

A quick look over some ODPs reveals that most of them suffer from a very confusing organisation of datasets. The first level of description 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 usually not connected.

As a result, the navigation, exploration and search within individual, but in particular also across ODPs, is significantly hampered. Thus, we present in the following the STODaP approach, whose intention is to facilitate the access to open data, improving inter- and intra-ODP datasets descriptors.

5.2 STODaP Architecture

In this section, we present our approach for cleaning up, enriching and reconciling metadata of open datasets, supported by software tools both at local and global contexts. The objective of this approach is to tackle the main problems identified by the metrics described in [Section 4.2](#), and thus to enhance open datasets description and link them through harmonic metadata.

[Figure 14](#)⁹ shows an overview of the proposed approach. Data publishers in charge of ODPs are offered tools for enhancing local tag curation and semantic lifting. These local tags are then connected to semantic tags hosted in a central server, which is automatically fed by data coming from ODPs. Data consumers have the option to retrieve data directly

⁹ Icons by [SimpleIcon](#) from [www.flaticon.com](#) are licensed under [CC BY 3.0](#).

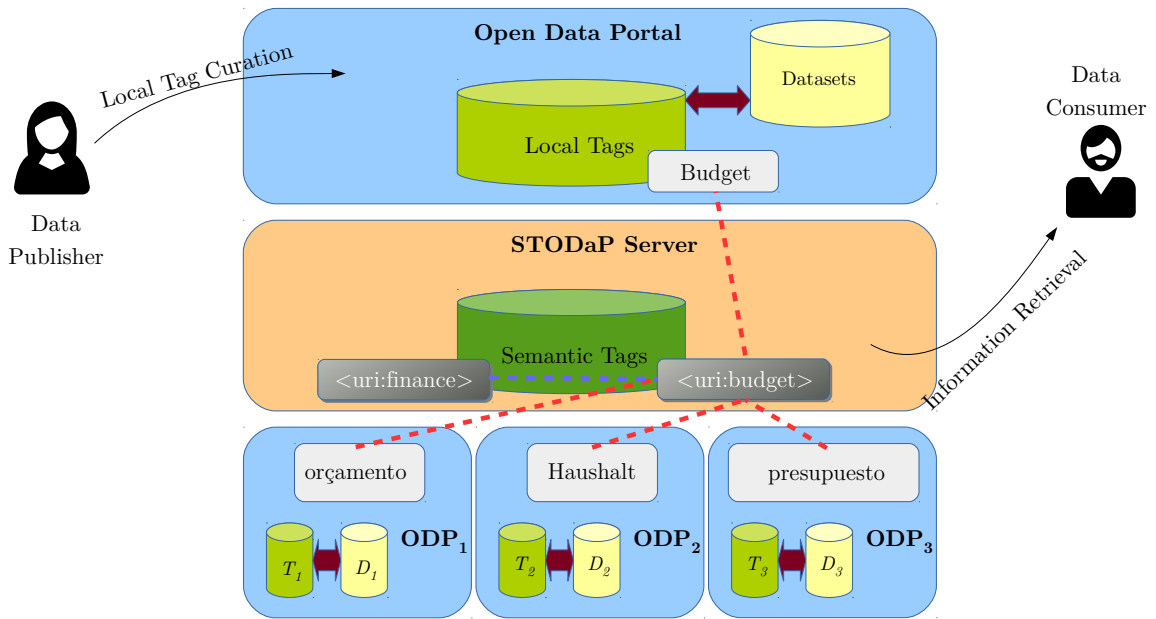


Figure 14 – Overview of the STODaP approach. Local tags are connected to a corresponding semantic tag within a central tag server. Data managers responsible for ODPs may use tools for local tag curation and semantic lifting of metadata.

from ODPs, or through references gathered from the central server. Semantic groups are also stored on the server and connected to local groups, but are omitted from the overview figure for simplicity.

This section is divided as follows: first, we present an overview of the STODaP architecture in [Subsection 5.2.1](#). The following subsections describe each element of this architecture, i.e., Open Data Portals ([Subsection 5.2.2](#)), external plugins ([Subsection 5.2.3](#)), local and global Processing steps ([Subsection 5.2.4](#) and [Subsection 5.2.5](#)), Semantic Metadata Repository ([Subsection 5.2.6](#)), STODaP vocabulary ([Subsection 5.2.7](#)), and external interfaces ([Subsection 5.2.8](#)).

5.2.1 Architecture Overview

[Figure 15](#) depicts the architecture of STODaP approach, showing all components and their connections. The approach is composed by the STODaP server, which hosts the most part of components, and the ODP extensions. STODaP receives as main input metadata of Open Data Portals, which are basically dataset descriptors such as title, description, tags, groups and others. These metadata are pre-processed individually for each portal at the *Local Metadata Processor* component, and stored at the *Metadata Repository*. The same metadata are also jointly processed, together with data coming from semantic knowledge bases from the Linked Open Data cloud, at the *Global Metadata Processor* component. This component outputs Semantic Tags and Groups, which are

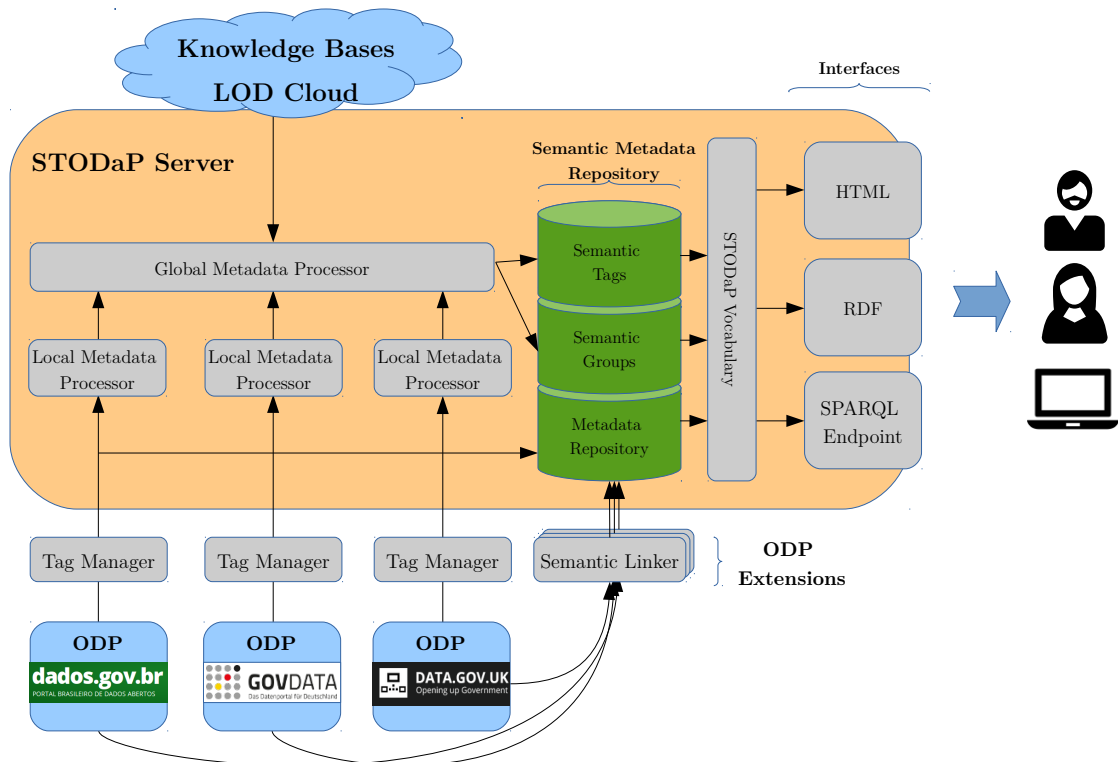


Figure 15 – Architecture of the STODaP approach. STODaP server receives input from ODPs and the LOD cloud (in blue). Grey blocks denote the components that will process metadata to reach the Semantic Metadata Repository, which will be accessed by the interface components in order to provide the output.

stored at the Semantic Metadata Repository and then coded using the STODaP vocabulary. Resultant dataset is made available for the general public through three types of interfaces: an HTML website where users can navigate manually, an RDF/XML interface which responds to machine requests searching for the resources URIs, and a SPARQL endpoint, which accepts queries and responds with JSON coded triples. In addition, the STODaP approach also envisages *ODP extensions* to enhance tag management and to link local tags with the server. The STODaP approach is independent of these components, and their operation is under responsibility of ODP administrators. In the following, each of these components is explained in details.

5.2.2 Open Data Portals

According to Colpaert et al. (2013), an Open Data Portal is “a collection of systems set up to make Open Data used and useful”. This definition sounds quite ambitious, since the great majority of ODPs are not a collection of systems, but of datasets. A formal definition of an ODP can be found in Umbrich, Neumaier and Polleres (2015), where all elements are mathematically described. In this section, we will define the only elements

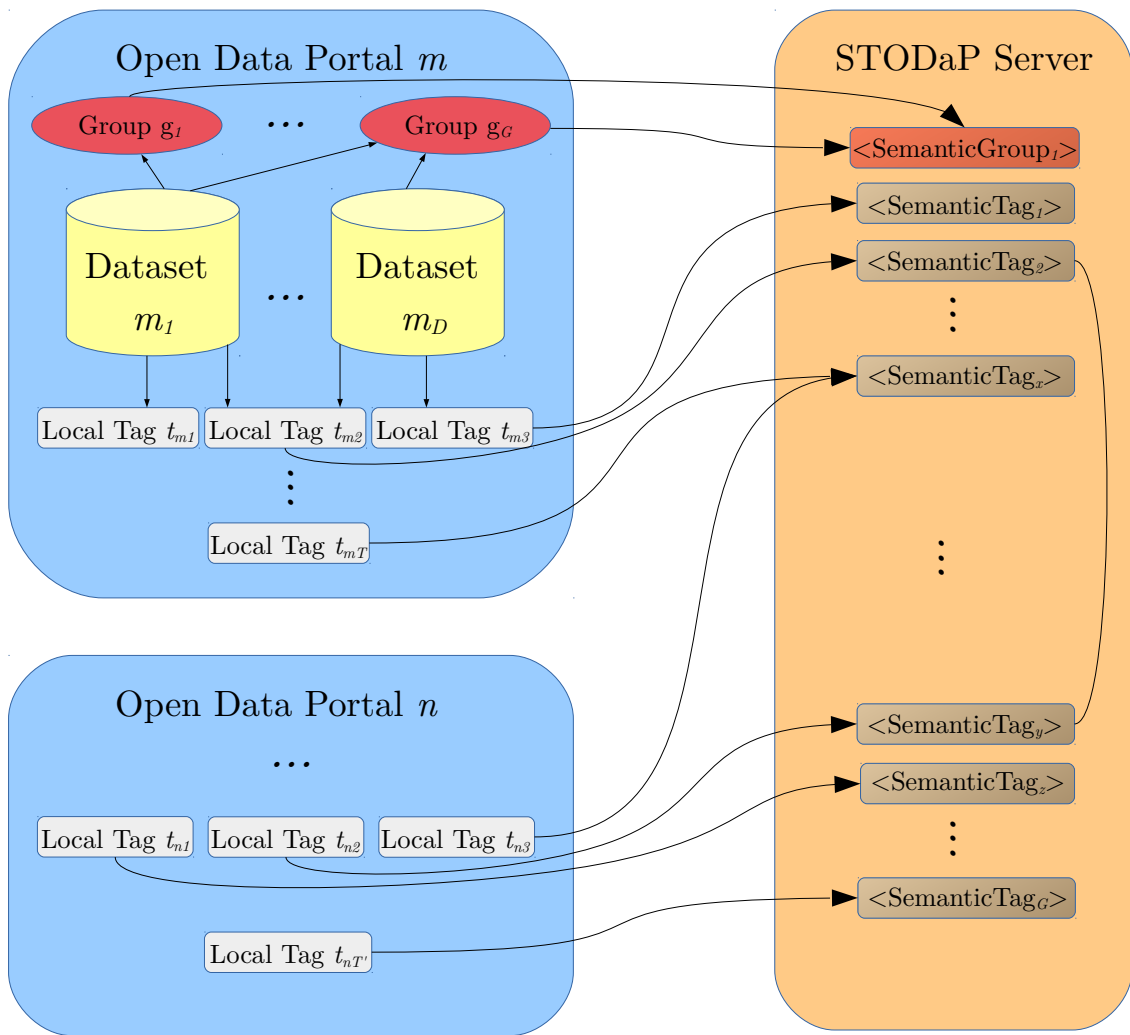


Figure 16 – Relevant elements of an Open Data Portal, and their connections to the STODaP server.

that are present in the context of this work, and their connections with the STODaP approach. This is shown in [Figure 16](#).

An *Open Data Portal*, in this context, is a collection of datasets, which can be owned by governments, NGOs, universities or other institutions. Open Data Portals are administrated by authorized users, who are in charge of uploading resources and filling metadata fields.

Datasets are containers that can hold one or more open data resources of several formats, including CSV, XLS, JSON, and even non-open ones, such as PDF. Most ODP metadata are associated to datasets, and the main ones are name, description, author, date, maintainer, and tags.

Groups are dataset containers, and are used normally to organise datasets by theme.

One group can hold several datasets, and one dataset can be associated to several groups (or none).

Tags are a metadata element defined by the string of characters that represents its name. Every distinct string represents a different tag, and every dataset tagged with tags represented by the same string can be sorted using this tag. Tags normally refer to a specific theme dealt by the dataset. However, it is quite frequent that tags represent temporal or geographical references, publishers or other kinds of informations. It must be emphasized that tags, in this context, are not related to a tagger, i.e., the person who tags. This happens in folksonomy contexts ([GRUBBER, 2007](#)), which, as previously discussed, is not our case. Tags are always associated to a context, in this case, an ODP. A dataset can be tagged with several (or no) tags, and a tag can be related to one or more datasets. At the CKAN platform, it is possible to have a tag not related to any dataset, however, we will not consider those ones. In our context, we will define these tags as *Local Tags*, in order to emphasize that they only exist in a single ODP context.

In order to illustrate the whole architecture, we also show the connections between Tags and *Semantic Tags*, and between Groups and *Semantic Groups*. Several local tags from different ODPs can be associated to a single semantic tag, which can also have semantic relationships with other semantic tags.

The STODaP approach also includes plugins to enhance metadata quality inside ODPs. Two of these add-ons will be presented in the following. It is important to highlight that these plugins are external to the STODaP server, and can only be installed and used by ODP administrators.

5.2.3 ODP Extensions

In this section, we describe two extensions that can act directly in the ODP in order to enhance the quality of metadata before sending it to the STODaP server. Although important, these components are optional to the approach.

Tag Manager

[Subsection 4.2.1](#) showed that ODPs suffer from low reuse of tags, and that there is a significant tags duplicity 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 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 clean-up and reconciliation of tags.

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

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 13](#).
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 considered.
3. Finally, we use Natural Language Toolkit ([BIRD; LOPER; KLEIN, 2009](#)) and the WordNet database 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 measures have to be computed for every pair of tags. 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.

Semantic Linker

After this cleaning procedure, we offer users the opportunity to link each local tag to a semantic correspondent at the STODaP server. The main idea is to enable not only the connection from STODaP server to ODPs, but also the other way around. The semantic tags plugin automatically suggests connections between local tags and semantic tags. Connections can also be done manually.

Linking local and semantic tags can bring several benefits for users navigating in ODPs, such as: better search options based on semantic tags; recommendation of similarly tagged datasets located in other portals; or taking advantage of the structure provided by the relationships between semantic tags, and consequently, between local tags.

–

In order to build the first version of the STODaP server, a metadata harvesting was driven through 87 ODPs. Almost 500.000 datasets were processed, including their metadata such as names, language, tags and groups they belong to. In the following subsection, we describe the procedures applied to the individual portals, in the process called Local Processing, which refers to the fact that it deals only with individual ODPs data.

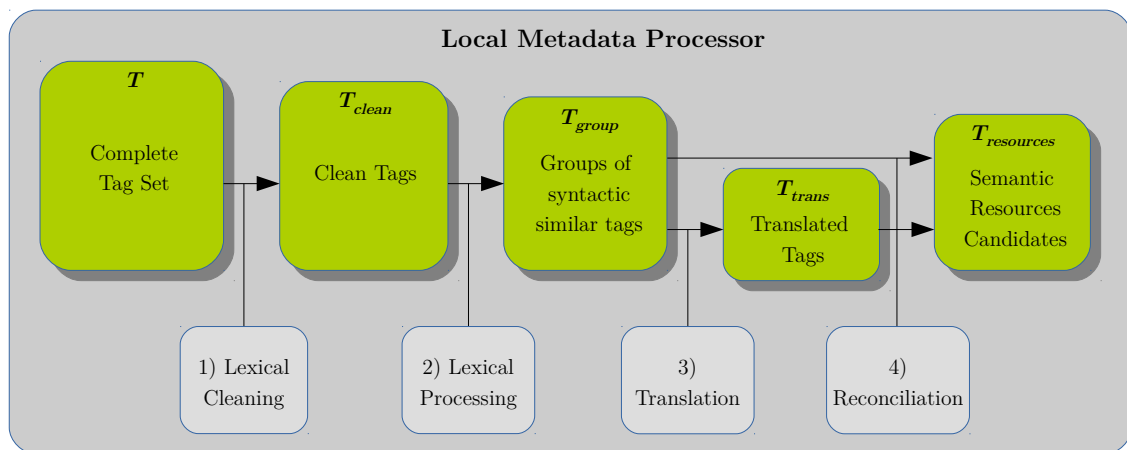


Figure 17 – Local tag processor. Green blocks represent the tag sets, that are transformed by the processes depicted by the grey blocks. The complete tag set suffers four transformations until reaching the last stage, when it is ready to be transformed into semantic tags.

5.2.4 Local Processor

The first processing step inside the STODaP server works over metadata of each ODP, and thus is called *Local Processing*. An overview of the procedure applied for each ODP is shown in Figure 17. Each green block represents a state of the processing phase. Grey blocks describe the transformations suffered by the tag set from one phase to another. The aim of local processing steps is to transform freely written tags into semantic resources that are candidates for representing the datasets they are associated. Processing starts with the complete tag set T , where each element is a local tag. Each transformation step is detailed below:

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

- smaller than 4 characters (tags such as `ac`, `aca`, `ad` would be discarded);
- composed by numbers and alphabetic characters (`50-year-rain-event`, `52-week` would be discarded);
- exclusively composed by uppercase characters (`APFO`, `EPSRC` would be discarded);
- not started by an alphanumeric character (`'Other' meteorological measurements`, `-10000` would be discarded);
- composed by more than 5 words (`Centre for Environmental Data and Recording`, `Countryside and Rights of Way Act` would be discarded); 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. This process 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 is 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 non-English tags from T_{group} was translated, resulting in a set of translations for each group. The new set achieved after this step 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 whole 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 uppercased 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 from edit-distance of 2 may also have different (or even opposed) meanings, such as

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	-, finanças, budget	{gemet:9366, eu-rovoc:825}, eionet:3194 eionet:1025

`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.

5.2.5 Global Processor

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

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

In order to accomplish this objective, we propose the process shown in [Figure 18](#), where each step is detailed in the following:

Significance Selection: We start the global processing with a set $\mathcal{T}_{resources}$, which is composed by $T_{resources}^p$ 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 these tags; (iii) the quality of semantic resources candidates. Selected resources form the $\mathcal{T}_{significant}$ set.

Semantic Processing: After this step, Semantic Tags will be derived by connecting resources from $\mathcal{T}_{significant}$ to local tags. Semantic Tags are entities defined by an URI, who have a main name in English, and point to local tags, which in turn connect to datasets located into Open Data Portals. The set containing all Semantic Tags is denoted by \mathcal{T} .

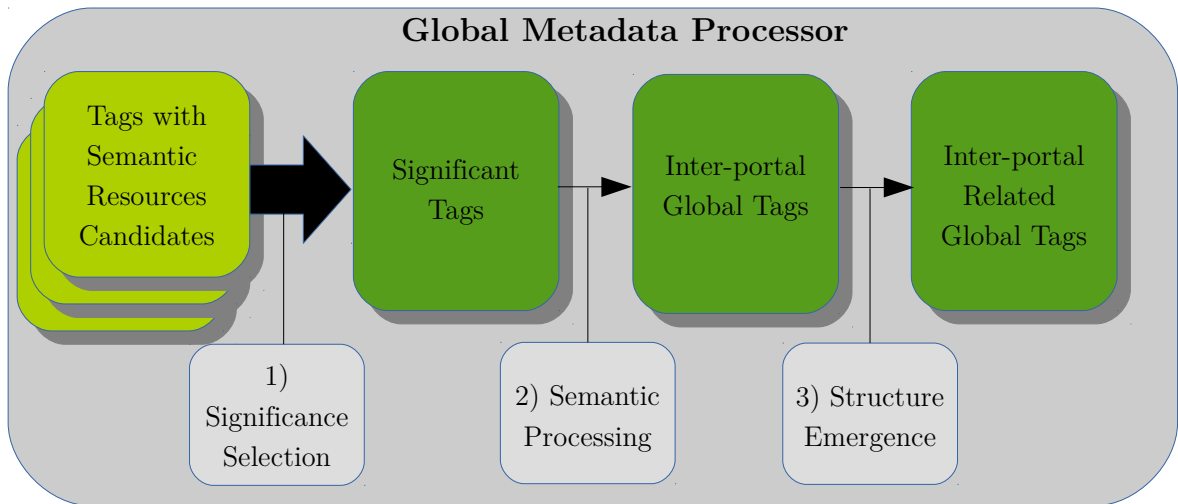


Figure 18 – Global metadata processor. Light green blocks represent the local tag sets resulting from the local processing of each ODP. Dark green blocks are the Semantic Tag sets, which are transformed by the processes depicted by the grey blocks. The global metadata processor outputs interlinked semantic tags.

Structure Emergence: Finally, relations between Semantic Tags in set \mathcal{T} will be searched on the ontologies they appear in order to give a structure to the Semantic Tag. The first strategy is to search for relations on the reconciled ontologies, and set this relation between the semantic tags. Thus, relations as `skos:related`, `skos:narrower` and `skos:broader` can be set.

At this point we notice the need of an upper classification scheme. The ODP model, as shown in [Figure 16](#), 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 organised in groups, it is possible to see a similar rationale. In [Table 11](#), we list the groups of 4 ODPs. Examining the table, it is clear that in the context of open government data portals, there are some context specific categories, but 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 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.

Table 11 – Examples of groups in some ODPs. Groups of Non-English portals are translated. Apart of a few context specific groups such as Multi-Year Plan and Municipal Chamber, the majority of groups fits generically to the Open Government Data context.

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

Finally, groups were reconciled with general-purpose ontologies. Particularly, the top concepts of Gemet Thesaurus¹⁰ fits for this purpose.

5.2.6 Semantic Metadata Repository

After the Global Processing steps, Semantic Tags and Groups are ready to be stored at the *Semantic Metadata Repository*, together with metadata originally collected from ODPs. Metadata are stored in a relational database, whose main models are

- Open Data Portal
- Dataset
- Tag
- Group
- Semantic Tag
- Semantic Group

¹⁰ Available at <http://www.eionet.europa.eu/gemet/>

5.2.7 STODaP Vocabulary

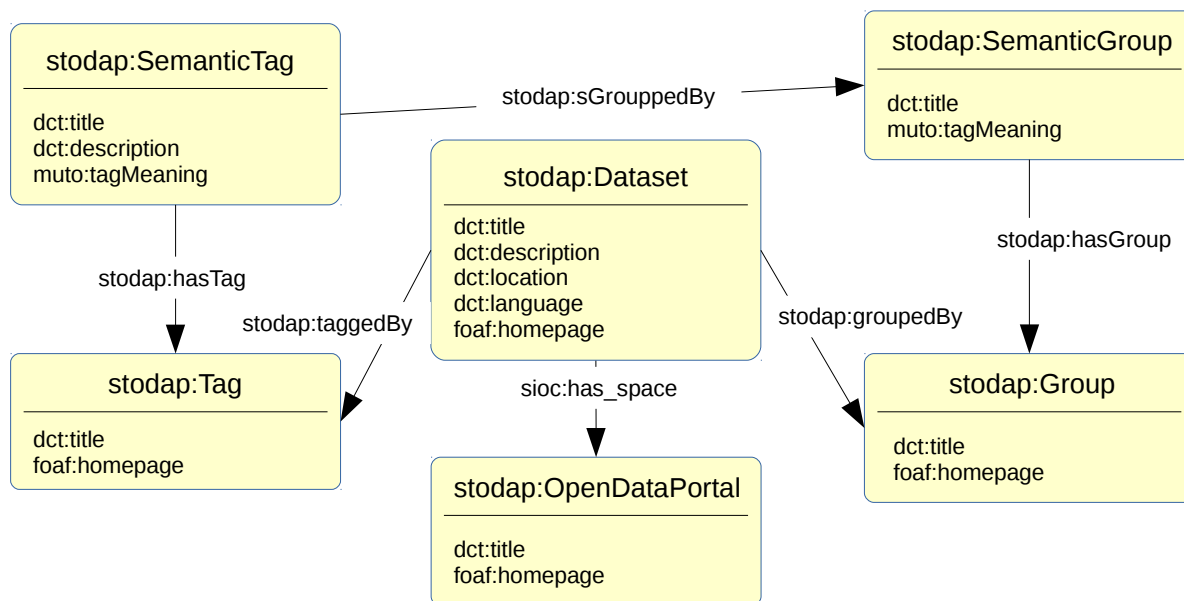


Figure 19 – Simplified schema of the STODaP vocabulary. Some elements are equivalent to other vocabularies and ontologies, such as SIOC, DCAT and MUTO.

In order to represent data in our approach, it was necessary to create a simple vocabulary. Figure 19 shows a simplified schema of the STODaP vocabulary. As shown in Section 4.1, several works describe ontologies and vocabularies related to our work. However, it was not possible to fit all entities of the STODaP model on existing ontologies. Specifically, DCAT¹¹ defines some important entities, such as `dcat:Dataset` and `dcat:Catalog`. They are defined as equivalent (`owl:equivalentClass`) to `stodap:Dataset` and `stodap:OpenDataPortal`, respectively. At DCAT, tags are represented as literals, which means that two tags with the same label do not differ. This is not the case at STODaP, where each tag is an entity. MUTO¹² tackles this issue with the class `muto:Tag`, equivalent to our `stodap:Tag`. On the semantic side, MUTO systematized the relation between a tag and a meaning with the `muto:tagMeaning` property. Thus, although some concepts were reused, `stodap:SemanticTag` and `stodap:SemanticGroup` needed to be defined. It must also be noted that MUTO works with a social concept of tagging, and thus defines a `muto:Tagging` class to enable relating actor to a tagging event, as described by Grubber (2007). Since the Open Government Data domain is not social, at least in the sense of tagging, this was not necessary in our case.

stodap:SemanticTag: A Semantic Tag is a super tag that groups open data portal tags and is connected to a semantic resource on the Linked Open Data Cloud.

¹¹ Available at <http://www.w3.org/ns/dcat>.

¹² Available at <http://muto.socialtagging.org/>.

stodap:SemanticGroup: A `stodap:SemanticGroup` is a super group of tags that groups open data portal groups, open data portal tags, and semantic tags. It is connected to a semantic resource on the Linked Open Data Cloud.

5.2.8 Interfaces

The last components of the STODaP architecture presented here are the Interfaces. Items in this component are designed to make data provided by the STODaP Server available for the external audience.

In order to respond to the various kinds of actors interacting with the STODaP server, 3 types of interface were designed.

HTML - Human browsable interface

The first interface is designed to offer ODP information for humans accessing the STODaP server. HTML browser interface provides several options for users willing to find open datasets. One option is to navigate through Semantic Tags. Each Semantic Tag points to related local tags, which in turn are linked to tagged datasets. Users are also presented to related Semantic Tags (broader, narrower or related). Following the same reasoning, it is also possible to navigate through Semantic Groups and their related Semantic Tags, Groups and Datasets.

Besides navigation, the server also offers a keyword search interface. In order to take advantage of the semantically enhanced metadata, a faceted search was designed. Search is made in two steps: (i) user inserts a keyword; (ii) resulting datasets are presented, and can be filtered by 5 different facets: (a) Semantic Tags (b) Semantic Groups (c) Language (d) ODP (e) Country.

RDF - Dereferenceable URIs

In order to be compatible with the fourth star of open data ([BERNERS-LEE, 2010](#)), each element of the STODaP approach - Semantic Tags, Semantic Groups, Tags, Groups, Datasets and Open Data Portals have their own Unique Resource Identifier (URI). This URI is also valid as URL, and can be accessed via the RDF interface, which responds with an RDF document containing the attributes of the referred element.

SPARQL Endpoint

The SPARQL endpoint provides direct access to the information stored in the Semantic Metadata Repository coded with the STODaP vocabulary. Queries might be manually inserted, but can also be input via an API. This requirement is important to make the STODaP server compatible with automatic SPARQL queries generators, such as ExConQuer¹³.

¹³ Available at <http://eis.iai.uni-bonn.de/Projects/ExConQuer.html>.

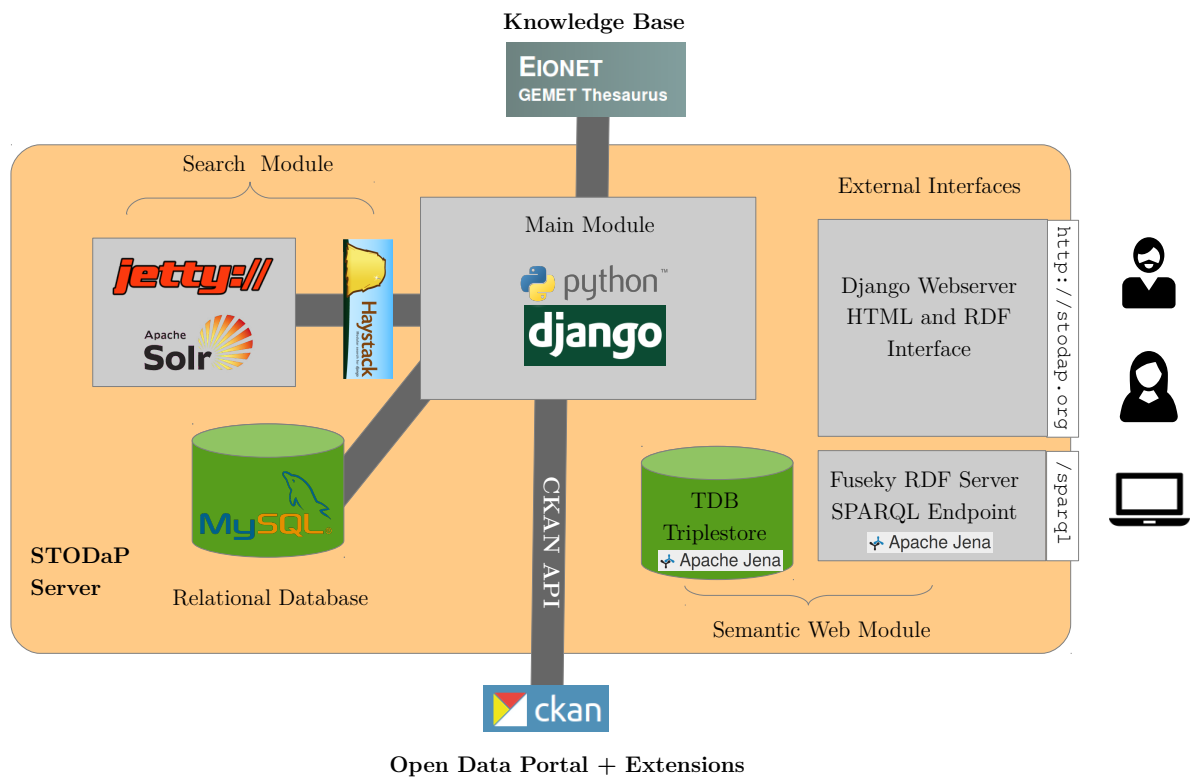


Figure 20 – Implementation architecture of the STODaP approach.

5.3 Implementation

In this section, we detail the technical choices related to the implementation of some elements of architecture presented in the previous section. We start with an overview about the implementation of STODaP server, detailing the software tools and integration strategies used, in [Subsection 5.3.1](#). A specific subsection is dedicated to the interface implementation, in [Subsection 5.3.2](#). Finally, we describe the implementation of the ODP extensions as CKAN plugins in [Subsection 5.3.3](#): (i) *CKAN Tag Manager* and (ii) *CKAN Semantic Tags*, which materialize the ideas reported in [Subsection 5.2.3](#).

5.3.1 Semantic Tags Server

The first version of the STODaP approach, presented in [Tygel et al. \(2016a\)](#), was implemented using *MediaWiki* and specially the *Semantic MediaWiki* extension ([KRÖTZSCH et al., 2007](#)). This extension turns the Wiki tool into a Semantic repository, facilitating the integration of objects into the Linked Open Data Cloud.

A second version of the STODaP approach was developed, and the need of a more complete search platform, not present in MediaWiki, was priority. Thus, the most appropriate technological choice was to build an interface from scratch, using a framework

that could be integrated with a search platform. The implementation architecture can be seen in [Figure 20](#).

As core framework, Django framework for Python Language was used. This framework offers a rapid prototyping environment, with database integration and a web server for development purposes. Django was integrated with a MySQL server, where the semantic metadata repository is stored in a relational database.

This database is indexed using Haystack Django plugin, which connects Django with an Apache Solr search platform. Searching design options such as weights, facets and keyword logics are defined in Django and transformed into an XML configuration file which is used by Solr. After this definition, Solr starts the indexing process, which enables the search mechanism.

In order to generate RDF triples, a Django2RDF converter was developed, based on the STODaP vocabulary. The converter reads Django models and generates RDF files for each class: Dataset, Tag, Group, Semantic Tag, Semantic Group and Open Data Portal. These files are uploaded into TDB Triplestore, which connects to the Fuseki RDF Server enabling a SPARQL endpoint.

5.3.2 Interfaces

Interfaces were implemented using Django Template language, which mixes HTML and a template syntax which allows basic logic operation (loops and conditions) and access to variables passed by the system. CSS and Javascript were also used to build the screens.

[Figure 21](#) shows the STODaP welcome screen. An introductory text is presented, and is illustrated by a simplified STODaP architecture. Some Semantic Tags and Groups are shown in order to motivate visitors.

An alphabetically ordered list of Semantic Tags is shown in [Figure 22](#), and a specific Semantic Tag page can be seen in [Figure 23](#). The faceted search interface is presented in [Figure 24](#), and the SPARQL endpoint is shown in [Figure 25](#). The complete implementation of STODaP server can be accessed at <http://stodap.org>.

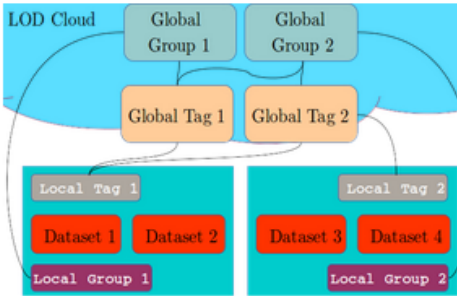
5.3.3 CKAN Plugins

CKAN offers an intuitive plugin development environment, which enables developers to modify or extend core functionalities of the system. The plugin architecture brings advantages both to core maintainers, that can keep their focus on the main functionalities of the platform, and for site administrators, who can keep their instances only with desired functionalities. Plugins can be easily installed by downloading the code and modifying one line at the configuration file. Communication between plugins and CKAN core is done

STODaP - Semantic Tags for Open Data Portals

[Home](#) || [Semantic Tags](#) || [Semantic Groups](#) || [Open Data Portals](#) || [Search](#) || [Vocabulary](#)

Welcome to the STODaP home!



The diagram illustrates the STODaP model. At the top, a blue 'LOD Cloud' contains 'Global Group 1' and 'Global Group 2'. Below it, 'Global Tag 1' and 'Global Tag 2' are shown. At the bottom, two 'Local Group' boxes (Local Group 1 and Local Group 2) contain 'Local Tag 1' and 'Local Tag 2' respectively. Each local tag is linked to specific datasets: Local Tag 1 to Dataset 1 and Dataset 2; Local Tag 2 to Dataset 3 and Dataset 4. Arrows indicate the relationships between these elements.

Welcome to the STODaP prototype. The Semantic Tags for Open Data Portals platform hosts Global Tags connected to several Open Data Portals (ODP) over the Web. Each Global Tag is semantically linked to resources over the Linked Open Data Cloud, and points to related datasets in ODPs.

The figure on the left explains the model used in this application. After analysing data from almost 90 Open Data Portals, we build a dataset of Global Groups and Global Tags. These are semantic elements that groups real portal tags, which in turn are linked to open datasets.

In the STODaP platform, you have pointers to over 400.000 open datasets, which are semantically connected through global tags and groups. The Global Tags are also linked by broader/narrow/related relationships, based on the Gemet Ontology.

Navigate through [Global Groups](#), [Global Tags](#) or [Search](#) directly for your desired open dataset!

Global Groups

[agriculture](#), [air](#), [animal husbandry](#), [biology](#), [building](#), [see more...](#)

Global Tags

[access to culture](#), [access to information](#), [accident](#), [accounting](#), [acid](#), [see more...](#)

GRECO/PPGI/UFRJ - EIS/University of Bonn

Figure 21 – STODaP welcome screen. An introduction text is presented, with a diagram showing the STODaP model. Some Semantic Tags and Groups are also displayed in order to motivate visitors.

via CKAN Api¹⁴. Plugin can also implement logic operations and create new interface templates.

In order to implement tag cleaning and semantic linking functionalities, two plugins were implemented, and their description is in the following:

CKAN Tag Manager Plugin

The CKAN Tag Manager Plugin offers an environment for tag curation directly inside the CKAN platform. It comprises basic functions such as deletion and editing of tags (not present in CKAN core), and advanced function aimed to enhance the quality of tags. In this sense, the plugin looks for:

- Very similar tags, differing only 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](#)) and the WordNet database.

¹⁴ Available at <http://docs.ckan.org/en/latest/api/index.html>.



Figure 22 – STODaP Semantic Tags. An alphabetically organised list is presented, in order to enable user navigation. In brackets, the number of datasets related to each tag.

In all those cases, user is offered the option of merging the respective pair of tags. Figure 26 shows a screenshot of the CKAN Tag Manager. The plugin was developed using CKAN API in Python language, and can be installed in any CKAN instance. The source code is available for download and contributions at <https://github.com/alantysel/ckanext-tagmanager>.

CKAN Semantic Tags Plugin

The CKAN Semantic Tags plugin implements the connection between a CKAN instance and the Semantic Tag Server. Each local tag can be associated to a semantic tag from the server. After the association, datasets linked with a local tag also point to the global server, as shown in Figure 27.

The plugin was developed using CKAN API in Python language, and can be installed in any CKAN instance. The source code is available for download and contributions at <https://github.com/alantysel/ckanext-semanticstags>.

5.3.4 Use and Maintenance of the STODaP server

After populating 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 at the server level.

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 server, a setup procedure is followed:

- Harvest tags, datasets and groups metadata using CKAN API;
- Drive the *Local Processing* described above;
- Reconcile tags with existing Semantic Tags;
- If reconciliation is not successful, search lexvo.org and try to create a new semantic tag;
- The same is applied for groups.

Updating a portal

ODPs are very dynamic, and have to be constantly updated. This procedure can be done on demand, and consists of:

- Harvest tags, datasets and groups metadata;
- Verify which metadata was changed, and for those, apply the procedure described above.

5.4 Quantitative Results

We describe in this section some quantitative results achieved with the STODaP approach. A deeper qualitative evaluation is described in the next chapter. At the global level, we analyse some aspects of the implementation. At the local level it is only possible to claim potential results, as we do not have access to the single ODPs.

5.4.1 STODaP Server

From a total of 291.805 Local Tags, we extracted 2142 Semantic Tags, all linked to the Gemet Thesaurus. The 1743 Local Groups resulted in 74 Semantic Groups, which are linked to the top concepts of the Gemet Thesaurus. Between Semantic Tags, we found 3314 relations, being 1355 typed as `skos:broader`, 1355 as `skos:narrower` and 604 as `skos:related`.

An important analysis is related to the number of local tags related to semantic tags. If we look at [Subsection 5.2.4](#), local tags may be discarded on *Local Processing*, if they do not meet the criteria, or on *Reconciliation* step, if an adequate semantic resource is not found for this tag. After the whole procedure, 23,850 local tags were associated to semantic tags, which represents only 8.22% of the corpus. However, because the most relevant tags were selected, 152,625 datasets are tagged by these local tags. This figure represents 32.44% of all datasets. Moreover, if we take into account that 172,157 datasets (or 36.59%) are not tagged by any tag, we can say that 51.15% of tagged datasets are referenced by a semantic tag.

Reasons why 48,85% are not linked to semantic tags may lie on metadata quality and on the proposed approach. Enhancing the semantic lifting procedure can certainly increase this percentage, however at some point still to be determined, lack of metadata quality will prevent further improvements.

5.4.2 Local Level

At the local level, the main potential achievements are at the tag curation process. As shown in [Figure 13](#), 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.

From the previous discussion, it is clear that enhancing metadata quality is fundamental to increase success on their semantic lifting. Strategies at the local are of paramount importance, but can only be performed by ODP administrators.

5.5 Conclusions

In this chapter, we presented an approach for metadata reconciliation within and among Open Data Portals. The main objective of this approach is to tackle the open data description problem, which was shown to be relevant in previous chapters. In [Chapter 3](#), one of the results of a participatory research based on Data Literacy courses pointed out to the difficulties of novice users in finding open datasets because of the lack of high quality descriptors. A literature based research in [Chapter 2](#) and [Chapter 4](#) also pointed to the description problem as relevant to the development of open data. On the analysis driven in previous chapter, we found that several portals share the same tags, showing that this specific metadata has a good potential to be a linking elements among datasets.

Converting tags into semantic identifiers was also shown to be a viable option, even though more sophisticated methods for semantic lifting are still 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 open datasets descriptors, and a global one, for connecting ODPs through semantic tags. The implementation of both shows that significant enhancements can be achieved both at individual ODPs and globally. In the next chapter, an evaluation of this approach will be presented.

STODaP - Semantic Tags for Open Data Portals

[Home](#) || [Semantic Tags](#) || [Semantic Groups](#) || [Open Data Portals](#) || [Search](#) || [Vocabulary](#)

[Home](#) > [Global Tags](#) > cadmium

[RDF](#)

cadmium

One of the toxic heavy metal which has caused deaths and permanent illnesses in a series of major pollution incidents around the world. Cadmium has no useful biological purpose. However, it has wide industrial applications. It has been used for decades in metal plating to prevent corrosion, in rechargeable batteries and as a pigment in certain plastics and paints. Special care is taken in the industrial smelting of ores and subsequent handling of cadmium, because occupational exposure is known to have caused heart, chest and kidney disorders. Environmental health problems have come from exposure to various sources of pollution.

URI: <http://www.eionet.europa.eu/gemet/concept/1100>

Global Groups

chemistry

Local Tags

Cadmium (2)

cadmium (15)

Cadmium (18)

cadmium (38)

cadmium (1)

Cadmium (1)

cadmium (2)

cadmium (7)

Related

heavy metal

Datasets

<http://data.gov.uk> (Cadmium)

[AFBI Soil Geochemical Map of total Cadmium for Northern Ireland \(Metadata\).](#)

The AFBI Soil Geochemical map of total cadmium (aqua regia digestion, mg/kg) for Northern Ireland is a 1km grid map, interpolated (using kriging) from over 6000 topsoil samples.

AFBI

Aqri Food and Biosciences Institute

Agricultural and Aquaculture Facilities

Agricultural and aquaculture facilities

Agriculture

Agriculture and Fishing

Aqua Regia

Cadmium

Geochemical

Geochemical Soil Map

Geochemistry

Geology

INSPIRE

Mapping

Metadata

NI

Northern Ireland

Soil

Topsoil

Topsoil Samples

agriculture

geology

soil

[AFBI Soil Geochemical Map of extractable Cadmium for Northern Ireland \(Metadata\).](#)

The AFBI Soil Geochemical map of 0.05M EDTA extractable cadmium (mg/kg) for Northern Ireland is a 1km grid map, interpolated (using kriging) from over 1000 topsoil samples.

AFBI

Aqri Food and Biosciences Institute

Agricultural and Aquaculture Facilities

Agricultural and aquaculture facilities

Agriculture

Agriculture and Fishing

Cadmium

Extractable

Geochemical

Geochemical Soil Map

Geochemistry

Geology

INSPIRE

Mapping

Metadata

NI

Northern Ireland

Soil

Topsoil

Topsoil Samples

agriculture

geology

soil

Figure 23 – Example of the cadmium Semantic Tag. The screen presents a description, the URI, the related Global Groups, Local Tags and Related Datasets

Faceted Search

Search:

3064 results

Filters

Click on filter elements to narrow your search.

Semantic Tags

- [budget](#) (961)
- [finances](#) (161)
- [city](#) (113)

Semantic Groups

- [economics, finance and work](#) (1098)
- [public administration](#) (541)
- [population](#) (164)

Language

- [en](#) (2616)
- [de](#) (140)
- [es](#) (121)

Portals

- <http://catalog.data.gov> (1209)
- <http://datahub.io/> (465)
- <http://open-data.europa.eu/data> (207)

Country

- [United States of America](#) (1209)
- [British Indian Ocean Territory](#) (503)
- [Undefined](#) (386)

http://data.gov.uk (Budget Management)

[Budget Management](#)

Budget Management [\(Homepage\)](#) [\(RDF\)](#)

http://datahub.io (Slovenian Budgets)

[Slovenian Budgets](#)

Slovenian Budgets. [\(Homepage\)](#) [\(RDF\)](#)

http://datahub.io (CERN Budget)

[CERN Budget](#)

#CERN Budget [\(Homepage\)](#) [\(RDF\)](#)

http://catalog.data.gov (Budget 2012- CIP)

[Budget 2012- CIP](#)

Capital Improvements budget, 2012. More at [\(Homepage\)](#) [\(RDF\)](#)

Figure 24 – STODaP faceted search. After inputting a search keyword, facets are presented – Semantic Tags, Semantig Groups, Country, Language and Portal – with the number of results in brackets. Users can narrow results by clicking on the facets.

```

6 PREFIX sioc: <http://rdfs.org/sioc/ns#>
7 PREFIX stodap: <http://stodap.org/>
8
9 SELECT ?dataset_name ?tag_name
10 WHERE
11 { ?dataset a stodap:Dataset;
12     |stodap:taggedBy ?tag;
13     |dct:title ?dataset_name.
14     ?tag dct:title ?tag_name.
15     <http://stodap.org/semantictag/3643.rdf> stodap:hasTag ?tag.
16 } {?dataset dct:location "Brazil".} UNION {?dataset dct:location "Uruguay".}
17 }
18
19

```

QUERY RESULTS

Table Raw Response

Showing 1 to 4 of 4 entries

	dataset_name	tag_name
1	"Orçamento Federal"	"orçamento"
2	"Crédito presupuestal asignado y ejecutado con aperturas"	"presupuesto"
3	"Órgãos Estaduais"	"Orçamento"
4	"Gastos do Poder Executivo"	"orçamento"

Showing 1 to 4 of 4 entries

Figure 25 – STODaP SPARQL endpoint. The query asks for datasets tagged with local tags related to the *Budget* semantic tag, which is represented by the URI <http://stodap.org/semantictag/3643/>. Results are shown on the bottom of the figure.

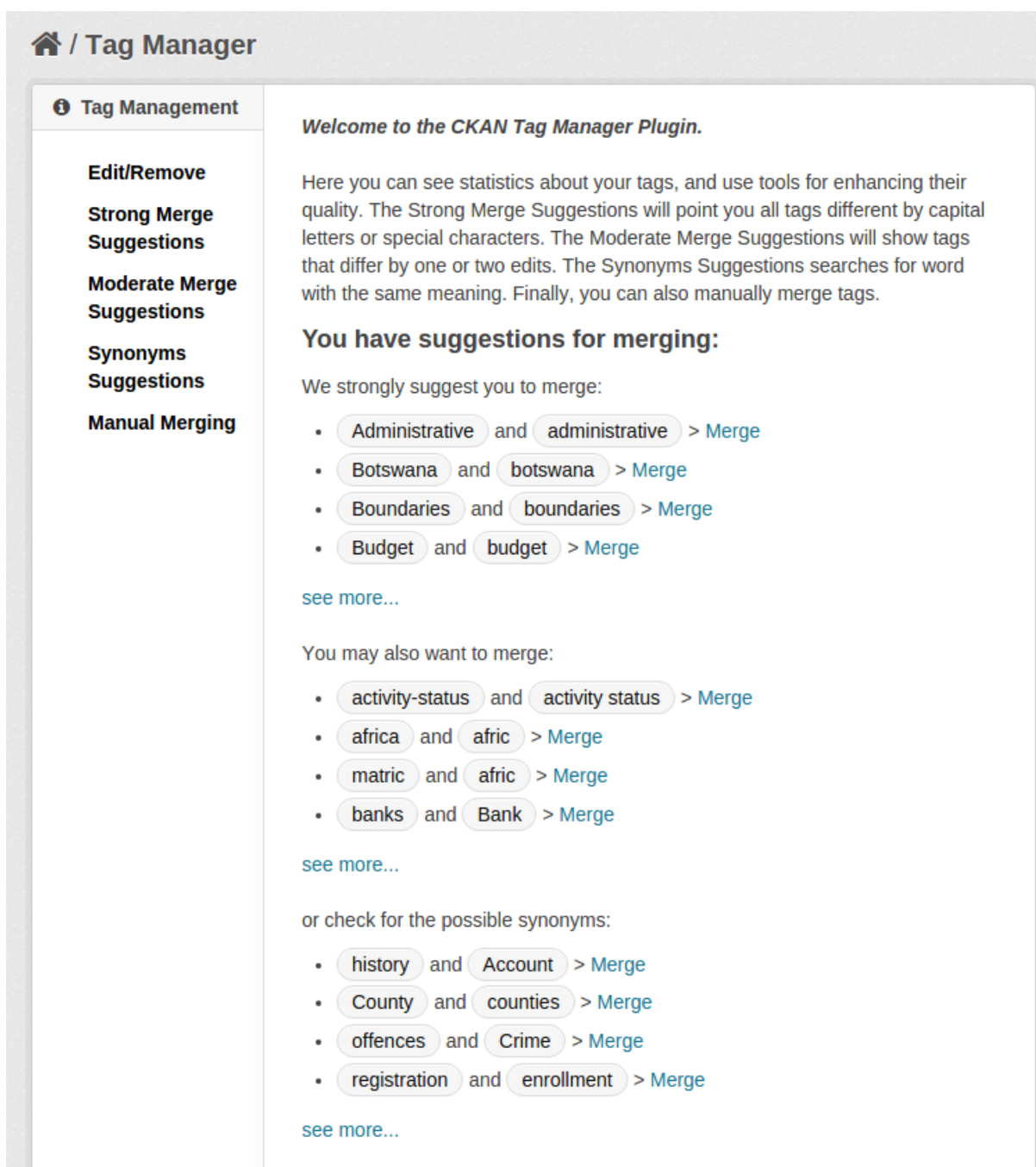


Figure 26 – Screenshot of the Tag Manager plugin, for tag curation in a CKAN instance. The plugin offers possibilities of manual and semi-automatic tag merging. The first block contains only valid suggestions, while the second block shows 2 false-positives. The synonym module also detected plurals. Tags in this example were extracted from the africaopendata.org portal.

The screenshot shows a web interface for a dataset. At the top, there is a navigation bar with four items: 'Conjunto de dados' (selected), 'Grupos', 'Fluxo de Atividades', and 'Relacionado'. Below this is the title 'Preços de Alimentos' and a subtitle 'Índice de preços de alimentos, medido pela FAO - Organização das Nações Unidas para alimentação.'. Underneath is the section 'Dados e recursos' with an 'XLS' icon, the title 'Food price index', and a description 'The FAO Food Price Index is a measure of the monthly change in international...'. To the right of the description is a blue button labeled 'Explorar'. At the bottom, there is a horizontal bar containing three tags: 'FAO', 'alimento', and 'Food'. The 'Food' tag is highlighted in red and has a mouse cursor pointing to it.

Figure 27 – Screenshot of the Semantig Tag plugin. The dataset is tagged with two tags, and one of them (`alimento`) is connected to the semantic tag `http://stodap.org/semantictag/3996/` through the `mut0:hasTag` property.

6 Evaluation

In the previous chapter, the 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 organisation problem.

In this chapter, an evaluation of the STODaP approach is described. STODaP was compared to other mechanisms on the task of searching for open datasets. We first present an overview of the evaluation concepts, followed by a theoretical background on search engine evaluation methodologies in [Section 6.2](#). Then, we show the experimental setup in [Section 6.3](#), the pre-evaluation procedure in [Section 6.4](#) and the evaluation itself, together with its results in [Section 6.5](#). Finally, some aspects of the evaluation are discussed in [Section 6.6](#) and conclusions are included in the final section.

6.1 Overview

In this section, we present an overview about the rationale followed in order to evaluate the STODaP approach. Limits of this process are also detailed.

First of all, it is important to state what exactly is going to be evaluated. As seen on the previous chapter, the STODaP approach consists of several components, including the STODaP server, whose input are ODP metadata, and the output are interfaces to access the semantically lifted metadata. This approach was implemented, as described in [Section 5.3](#).

Semantic lifting of metadata is done not for the sake of doing, but to facilitate the task of users wanting to find open datasets on the web. Thus, our aim in this chapter is to evaluate if the searchability of open datasets was enhanced with the use of STODaP server.

One of the conclusions of [Chapter 3](#), which can be seen in [Table 23](#), is that the way open data is published and described still imposes several barriers for some users. Some comments by the participants of the Data Literacy course attest that “Finding data in the web is hard”, “Open data portals are complicated” and “Data organisation is confused”. Thus, by evaluating our proposed approach as an open dataset search engine, we are able to have clues about the problems pointed out during the participatory research.

This rationale, however, imposes some limitations to our evaluation process. First of all, for several reasons that will be discussed in the next chapter, it was not possible for the participants of the Data Literacy course to take part in this evaluation. Moreover, this kind of evaluation results that some elements of the STODaP approach are not being

assessed. For example, ODP extensions are not being proofed, since we were not able to install their implementation on real world ODPs. Assessing the final output of the STODaP approach does not mean that we have evaluated the quality of semantic lifting, the amount of relations between tags, or in which extent tags really describe data inside open datasets.

Although it can be said that this evaluation is incomplete in some senses, we decided to favour the assessment of the initial purpose of this thesis, i.e., if the semantic lifting of metadata would benefit the searchability of open datasets.

Considering these limitations, the methodology for deriving this evaluation is the following:

- First, a literature review about evaluation of search mechanisms is driven;
- Then, the steps of the evaluation procedure are proposed and related support materials, such as text and evaluation software are implemented;
- In the sequence, the procedure is applied to a pilot group, and reviewed after the contributions of this group;
- The procedure is then applied to the main group of participants; and
- Results of this evaluation are analysed and discussed.

Each of these steps are reported in the subsequent sections.

6.2 Literature Review

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 and 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 rationale, the Normalized User Effectiveness (NUE) evaluates the relation between relevant documents retrieved by novices and experts, proportional to NT. The 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 the opposite direction, [Xu and Mease \(2009\)](#) defend the use of task completion time as a robust measure to assess in which extent a 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. For tasks T1 and T2, we would have half of the users performing performing T1 with method A and T2 with method B, and half of them performing T1-B and T2-A.

In a survey dealing specifically with faceted search, [Wei et al. \(2013\)](#) present 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., ability of users, and factors 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, the 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 to precision and (ii) *Ability to retrieve top ranked pages*, as counterpart measure to 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.

6.3 Experimental Setup

In this section, we describe in details our experimental setup. First we discuss the goals of the evaluation and the metrics used to quantify the experiments. Then, search engines used to compare the performance of the STODaP server are presented, followed by the participants profile, questions, procedure and validation of results.

As explained in [Section 6.1](#), the procedure was adjusted after the pre-evaluation. Thus, the version presented here is the modified one, applied in [Section 6.5](#).

6.3.1 Goals

First of all, we define the evaluation goals:

- **G1:** When searching for open datasets, how does the STODaP server compares to other data-specific and general search engines?
- **G2:** Is the STODaP server an useful tool for searching open datasets?

G1 will be answered through objective assessments. Metrics in this case will be Task Completion Time (TCT), and Precision. Both will be defined in [Subsection 6.3.6](#).

The second goal (G2) relies on subjective evaluation of participants. This goal will be assessed via a questionnaire presented after the experiment ([Table 13](#)), and contains both a question about absolute user satisfaction, and another related to satisfaction in relation to other methods.

6.3.2 Comparative Assessment

In order to make a relative evaluation of STODaP server, we looked for other systems developed to facilitate the search for open datasets.

The first option could be CKAN tool itself, as far as it is able to store open data metadata, and offers a search mechanism for it. Datahub.io¹ is an example of CKAN based general purpose ODP. Using CKAN plugins such as ckan-harvest², it is possible to harvest datasets from other CKAN ODPs and include their metadata in its own base. Datahub.io also offers users a home for its data: general public is able to create an account and upload data in there. However, for this reason, comparing Datahub.io with STODaP would no be appropriate, because the nature of data is different in both tools.

Dataopen.eu³ is a full text search engine for the European Union. Their home-page clearly warns visitors: “We are not like CKAN: we index the contents of open data, not only metadata.”. This approach is quite interesting, and inspires the future steps of STODaP development. However, being restricted to the European Union, it is also inappropriate for our comparison purposes.

Finally, a tool was found with a similar purpose and scope as STODaP:

Exversion: Exversion Data Search Engine⁴ is a platform for groups wanting to control, collaborate and share their data. It can be used to host both closed and open data, and is able to connect to data editors via a REST API. Recently, Exversion launched a Data Search Engine, which indexes open data portals by scrapping its content, and provides a unified search interface to them ([EXVERSION, 2015](#)). Exversion has a smaller base (60

¹ Available at <http://datahub.io/>.

² Available at <https://github.com/ckan/ckanext-harvest/>.

³ Available at <http://dataopen.eu/>.

⁴ Available at <https://www.exversion.com/search/>

ODPs in July/2016) in relation to STODaP, but has the advantage of being able to index ODPs not using CKAN (Socrata, GeoNode, and others). Exversion does not include any semantic processing.

As observed in [Subsection 3.4.2](#), the first impulse of users when searching for open datasets is to use generic web search engines such as Google, DuckDuckGo, Yahoo, and others. Thus, we included a free search to let users choose their preferred search engine:

Free: Although not specialized in open datasets, generic search engines have a bigger processing power and several artificial intelligence algorithms that help users finding exactly what they want. As these algorithms are closed, it is not possible to affirm in which extent semantics is used in generic search engines. However, the use of knowledge graphs, synonyms identification, spell-checker and named entity extraction can be easily identified, besides tracking mechanisms as past searches, browsing history and IP location ([BHATTACHARYA, 2014](#)).

6.3.3 Participants

The aim of STODaP is to facilitate access to open datasets 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 participants can at least imagine a potential use of data. English knowledge is also necessary, because labels of semantic tags and groups are still only in English language.

6.3.4 Questions

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

The evaluation questions were selected based on: (i) topic relevance of datasets on the open data community, based on criteria defined by Open Data Index⁵ (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 5.2](#). Broader conclusion would require evaluations of larger scales, which are over the scope of this thesis. Defined questions are:

- **Q1:** Find open datasets about **water quality** on **7 different rivers outside Europe**.

⁵ Available at <http://index.okfn.org/>.

- **Q2:** Find open datasets containing **2015 budget data** from locations in **5 different countries**.
- **Q3:** Find open datasets containing **procurement information** in **3 different languages**.

The number of questions was chosen in order to balance diversity, i.e., choosing as many topics as possible to enlarge experiment coverage, but also to guarantee that it is viable in terms of time, considering that all participants would answer all questions (using different search methods).

6.3.5 Procedure

The following procedure was driven during the evaluation process:

- The main idea of the project is presented, followed by an explanation about the evaluation itself;
- Participants fill the entry-questionnaire, shown in [Table 12](#);
- After finishing the form, three tasks are sequentially presented. Each task is a combination of a question Q and a search method M. Combinations for each participant are chosen in order to guarantee that each Q-M combination has approximately the same number of answers.
- For each task, participants are instructed to find the specified datasets and paste their URLs in the appropriated fields. A screenshot of the evaluation tool is shown in [Figure 28](#).
- The time taken to complete each task is automatically calculated.
- The evaluation questionnaire shown in [Table 13](#) is presented to the participants.
- Afterwards, each answer is manually checked and classified as correct or incorrect.

Table 12 – Entry questionnaire.

ID	Question	Range
Age	How old are you?	> 0
Internet	How often do you use internet?	1 - once a week; 5 - everyday
Data	How often do you use structured data in your study/work? (e.g. spreadsheets, charts, statistics, ...)	1 - never; 5 - always
Open Data	How often do you use open data in your study/work?	1 - never; 5 - always
English	What is your English proficiency level?	1 - low ; 5 - high

STODaP - Semantic Tags for Open Data Portals

Task 3

Find open datasets containing **2015 budget information** at any administrative level from **5 different countries**.

Use the [STODaP server](#) to complete the task. Please open the [STODaP](#) search engine in another tab, and **do not close this tab**. Use the fields bellow to paste the datasets URL found. We just need an URL pointing to a page where the dataset can be downloaded.

If needed, you can always use auxiliary tools as translators and information sources like Wikipedia to help you completing the task. The dataset search should be done as specified above.

Figure 28 – Evaluation framework. Figure shows the presentation of one task, which in this case is represented by Q2 (Budget) and STODaP search method. After the question and the search method are explained, five text fields are presented for participants to fill in with open datasets URLs. A notice on the bottom clarifies that auxiliary tools can be used.

6.3.6 Metrics

With the aim of assessing G1, and based on the literature review, we define two objective metrics to be used in this evaluation:

Task Completion Time (TCT): normally, TCT would denote the total time taken to finish a task. In our case, a task is defined by a question and a search method, and the three questions defined for this evaluation require a different number of datasets to be searched for. Thus we define TCT as *the average amount of seconds that a participant takes to find a specified open dataset within a task*. A task is finished when the participant clicks the “Finish” button, and the evaluation framework automatically calculates the time T taken for this task. To achieve TCT, we divide T by the number N of required

Table 13 – Evaluation questionnaire.

ID	Question	Range
Absolute Satisfaction	Do you think STODaP is a useful tool for finding data on the web?	1 - not useful; 5 - very useful
Relative Satisfaction	How easy is it was get the data you need using the STODaP in comparison with the other methods?	1 - harder; 5 - easier
Comments	If you have additional comments or suggestions, please write it here:	Free text

datasets to achieve TCT:

$$\text{TCT} = \frac{T}{N}, \quad (6.1)$$

where T is the amount of seconds a subject takes to finish a task, and N is the number of datasets required be each task.

Validation methods will be used to guarantee that tasks not performed adequately are not considered as valid. Normalized TCT, as proposed by [Cheng, Hu and Heidorn \(2010\)](#), was also considered. However, the structure to guarantee a reasonable number of experts was not available.

Precision: Each question requires participants to find N datasets according to a defined specification. Thus, we define Precision as the number of true positives, i.e., returned datasets truly corresponding to the question definition, divided by the sum of true positives and false positives (N):

$$P = \frac{a_v}{N} 100, \quad (6.2)$$

where N is the number of required answers for each task, and a_v is the number of correct answers given by a subject.

Normally, precision metric is combined with recall: while the first measures the number of true positives in relation to the responded universe (true positives + false negatives), recall assesses the correct answer in relation to all relevant elements. In our case, calculating recall is not viable, since it would require to estimate the number of relevant elements over more than 400.000 datasets, in case of STODaP server, and an unknown quantity in case of generic search engines.

6.3.7 Validation

Each entry-questionnaire is analysed in order to determine if it is valid to our evaluation, in terms of internet experience. Additionally, participants are only validated if they complete all tasks and the evaluation questionnaire. Individual answers are checked

in order to confirm if the dataset links provided are really valid answers according to the assigned question. Moreover, if precision is smaller than 33.3%, or if TCT is higher than 1000 s, a task is not considered valid.

6.4 Pre-Evaluation

In order to test and adjust our evaluation setup, we ran the process described above with a group of seven students of an Information Retrieval graduate course, at the Federal University of Rio de Janeiro, on the 12th of May 2016. The results of this evaluation round can be seen in [Table 14](#) and [Table 15](#). Although the main target of this pre-evaluation process was to assess the evaluation procedure (and not the STODaP server), it is useful to look at the results to have the first impressions.

Table 14 – Answers to the entry and evaluations questionnaires. Columns correspond to the entry-questionnaire, applied before the evaluation, and the evaluation one, filled afterwards. Full question text can be seen in [Table 12](#) and [Table 13](#). Task Completion Time (TCT) is the average number of seconds taken to finish the task. Precision is the percentage of answers considered valid, over a total of 15.

	Age	Internet	Data	Open Data	Absolute Satisfaction	Relative Satisfaction	TCT	Precision (%)
1	27	5	3	1	4	2	295.7 +/- 98.8	100
2	23	5	4	3	5	3	833.3 +/- 151.8	80
3	27	5	5	5	2	2	560.0 +/- 103.2	33
4	23	5	4	3	5	4	845.0 +/- 523.9	73
5	26	5	3	2	5	5	625.3 +/- 269.5	67
6	29	5	5	3	4	4	527.0 +/- 287.5	100
7	22	5	4	1	5	5	351.0 +/- 121.2	80

[Table 14](#) shows the answers of each participant to the questions both before and after running the evaluation, together with its average task completion time and precision. As expected, all participants are frequent internet users. Use of data in daily work or study is also high, but only one declared himself an open data expert. This participant was the only who did not considered STODaP an useful tool for finding data on the web. Four out of seven considered that completing the tasks with STODaP was easier than with other methods. The average Task Completion Time had a huge variation, with a minimum of 295 seconds and a maximum of 845 seconds. Through a manual procedure, each answer was verified in order to check if it really corresponded to the given task. The verification

Table 15 – Task Completion Time of the pre-evaluation test, in seconds. Each cell contains the number of seconds that one or more participants took to complete the task with the correspondent search method.

Questions / Search Methods	Q1: Water Quality	Q2: Budget information	Q3: Procurement	Average and Standard Deviation	Accepted Answers (%)
Exversion	723, 884, 468	235, 382	558, 518	538.3 +/- 198.5	78
STODaP	435, 493, 460	397, 184, 517, 1048	-	504.9 +/- 244.0	83
Free	1580	702	401, 217, 180, 1001, 729	687.1 +/- 456.3	63
Average and Standard Deviation	720.4 +/- 383.7	495.0 +/- 276.7	514.9 +/- 266.7		
Accepted Answers (%)	76	80	71		

was not strict, and answers were discarded only if they were clearly wrong, or blank. No significant correlation was verified between TCT and precision.

Table 15 shows the Task Completion Times for each task. We can easily notice that choosing a random generator for attributing question / search methods combination was a mistake, specially with a small number of rounds. There were 36 possibilities (6 orderings for questions \times 6 orderings for search methods) but only 7 rounds were used. Thus, the number of combinations between questions and search methods ended up quite unbalanced, and there was no combination of STODaP search method with the procurement task. Average TCT for tasks and search methods were also calculated.

After running the procedure, a conversation round was driven with participants in order to get insights both about the tool and the evaluation procedure. The following suggestions and comments were made:

a) Regarding the evaluations interface:

- Alert that search engines should be opened in another tab;
- Write the questions more clearly and specific (e.g., asking budget 2015 may include 2014-16?);
- On the final questions, positive answers to the system were at the left side, which is not usual and confused some participants;
- State more clearly that only the link to the dataset should be answered;
- Make it clear that users are allowed to use auxiliary tools such as translators or

Wikipedia in order to better perform the tasks;

- State more clearly that users should only look to the dataset title, and there is no need to open it; and
- Explain that only some specific portals are indexed, not all open datasets in the world.

b) Regarding the evaluation procedure:

- On the evaluation questionnaire, ask the English proficiency and other languages, and ask if the English language hampered the performance; and
- There were too few questions, and thus learning curve was not evaluated.

c) Regarding the STODaP faceted search interface:

- There should be an explanation about faceted search and how does it work;
- Regarding the Portal facets, it was suggested to write portals name instead of URLs, when this metadata is available;
- One participant reported that he took a while to realize the language facet. After seeing it, question Q3 could be very quickly answered; and
- It was suggested to include the possibility of making the query broader by including facets with OR.

d) There were some positive comments:

- It was noted that results presented by STODaP had a higher quality in relation to Exversion, mainly because the latter automatically uses an OR logic between two terms. This results in many unwanted outputs;
- The possibility of searching English keywords and getting multi-language results was also positively mentioned, because no other tool presents such feature; and
- The STODaP interface, in its search results, exhibits datasets with all its tags. This was positively noted, because it helps to decide quicker if a dataset is of interest or not.

Analysing participants while they were completing their tasks has also shown some new perspectives. It was noticed that some participants tried to look deep at datasets in order to verify if they met the task criteria. It should be more clearly stressed in the explanation that this is not necessary, since our objective is only to find datasets, and not to verify their quality. Some participants tried to use analytic tools such as Google Public Data. Its focus is rather on analysing (open) datasets than on making them available for download in machine readable formats. Thus, for our intentions, this is not considered open data and it should also be stated in the explanation.

After considering the above mentioned comments, the procedure was enhanced and applied to the main group. Results are described below.

6.5 Evaluation

In this section, we apply the evaluation procedure described in [Subsection 6.3.5](#) and present the results achieved.

6.5.1 Participants Profile

The experiment was completed by three different types of participants. The first participants were first year university students attending a class on the topic Introduction to Information Systems, at the Federal University of Rio de Janeiro, in Brazil. The second group of participants was formed by Semantic Web researchers in Bonn (Germany) and Rio de Janeiro (Brazil), while the third one was composed by members of a discussion group of open data practitioners in Brazil. An entry-questionnaire was filled by the participants, whose answers are summarized in [Table 16](#). Participation was not mandatory and non identified, and there was no reward for participants.

The average age of the 37 participants was 25.9 years. Although all of them use internet every day, direct experience with data processing is average, as well as with open data. As our tool is developed for non-experts, this sample is adequate to the experiment.

6.5.2 Task Completion Time Analysis

Although it may look simple to assess this metric as detailed in [Subsection 6.3.6](#), some practical questions arose while analysing data.

As detailed in [Subsection 6.3.7](#), in order to remove disturbing samples we considered only answers with precision higher than 33% and TCT lower than 1000 seconds. The first case is aimed to remove participants who gave up without trying to complete the task, or who did not understand the task. Regarding the second case, since some evaluations were

Table 16 – STODaP evaluation - summary of participants profile. Value represent the average answer to the related question shown in [Table 12](#)

Question	Average ($n = 34$)
Age	25.7
Internet	5
Data	3.3
Open Data	2.7
English	4.3

made online, we considered that participants who took more than 1000 seconds to find a dataset actually gave up and left the evaluation tool open.

Another issue is related to the computation of wrong and blank answers. If a subject answers a question with a dataset that do not correspond to what was asked, there are two options: either an effort was taken, but the question was misunderstood, or no effort was taken at all and a random answer was given⁶. The same reasoning applies to blank answers, but in this case the probability that no effort was taken looks higher. After these considerations, the issue remains: if a question requires 7 datasets, and 3 were correctly answered and 4 were left blank (or were wrong), is it fair to divide T by $N = 7$? Or should we divide it by 3, considering that effort was put only on those answers?

In order to evaluate the impact of these considerations, we define:

$$\text{TCT}_{nb} = \frac{T}{N_{nb}}, \quad (6.3)$$

where T is the amount of seconds a subject takes to finish a task, and $N_{nb} \leq N$ is the number of not blank answers, and

$$\text{TCT}_c = \frac{T}{N_c}, \quad (6.4)$$

where T is the amount of seconds a subject takes to finish a task, and $N_c \leq N$ is the number of datasets correctly answered. Note that $\text{TCT} \leq \text{TCT}_{nb} \leq \text{TCT}_c$. [Table 17](#), [Table 18](#) and [Table 19](#) present the results for TCT, TCT_{nb} and TCT_c , respectively.

Table 17 – Evaluation Results - TCT. Table presents TCT median and standard deviation for each method and question. In brackets, the number of considered samples.

	Q1: Water Quality	Q2: Budget	Q3: Procurement	Aggregate
Exversion	108 ± 92 (63)	121 ± 94 (65)	78 ± 72 (33)	108 ± 87 (161)
STODaP	60 ± 49 (77)	63 ± 170 (50)	67 ± 80 (42)	60 ± 109 (169)
Free	60 ± 126 (70)	44 ± 37 (50)	100 ± 96 (24)	70 ± 99 (144)
Aggregate	68 ± 95 (210)	60 ± 117 (165)	78 ± 84 (99)	

[Table 17](#) shows the general results for TCT. On each cell, values represent the median, standard deviation and the number of samples in brackets. Columns represent the different questions, and rows, the search methods. The last row presents the aggregate results for questions, and the last column, for search methods. STODaP presents the lowest aggregate TCT median (60 s), followed by free search (70 s, or 17% higher) and Exversion (108 s, or 80% higher). Looking at each question individually, it can be seen

⁶ In order to illustrate this case, one user answered “www.google.com” when asked for procurement datasets.

Table 18 – Evaluation Results - TCT_{nb} . Table presents TCT_{nb} median and standard deviation for each method and question. In brackets, the number of considered samples.

	Q1: Water Quality	Q2: Budget	Q3: Procurement	Aggregate
Exversion	127 ± 112 (52)	153 ± 120 (56)	111 ± 128 (28)	136 ± 121 (136)
STODaP	60 ± 49 (77)	63 ± 170 (49)	67 ± 80 (42)	60 ± 109 (168)
Free	64 ± 146 (62)	44 ± 37 (50)	100 ± 96 (24)	70 ± 109 (136)
Aggregate	74 ± 114 (191)	60 ± 129 (155)	83 ± 107 (94)	

Table 19 – Evaluation Results - TCT_c . Table presents TCT_c median and standard deviation for each method and question. In brackets, the number of considered samples.

	Q1: Water Quality	Q2: Budget	Q3: Procurement	Aggregate
Exversion	127 ± 112 (51)	153 ± 120 (55)	166 ± 126 (24)	141 ± 121 (130)
STODaP	60 ± 54 (75)	86 ± 168 (44)	67 ± 80 (42)	65 ± 110 (161)
Free	69 ± 146 (58)	55 ± 49 (43)	110 ± 339 (20)	87 ± 215 (121)
Aggregate	74 ± 114 (184)	89 ± 127 (142)	85 ± 200 (86)	

that for Q2 (Budget data), free search was faster than STODaP, and for Q1 both were equivalent. A possible explanation for the budget case is the high availability of this kind of data over the web. As shortly explained in [Section 2.5](#), and in details in [Tygel et al. \(2016a\)](#), releasing budget data is of crucial importance, and this topic is being prioritized by most of the countries. Thus, the high availability enables web search engines to deliver faster and more accurate results.

Regarding Q2, a possible explanation can be found on [Table 18](#) and [Table 19](#). Since Q2 required 7 answers, for this question there was a higher rate of blank answers both on Exversion and free search methods. Thus, it can be seen in the first column of [Table 18](#) that, while STODaP remains with the same value as in the previous table, the other search methods increased their TCT_{nb} because in those cases, $N_{nb} < N$. Interestingly, free search presented blank answers only for Q1. Q2 and Q3 have the same values in [Table 17](#) and [Table 18](#). The aggregate result also remained the same.

[Table 19](#) presents results related to TCT_c . In this case, STODaP advantage is even higher in relation to free search (34% higher) and Exversion (117% higher). However, in this scenario, free search is still faster for Q2. Results in this table are strongly related to the precision of the answers, which is discussed in the following section.

A last aspect must also be taken into account: standard deviation. As calculated here, this measure represents the average deviation of each value from the mean value.

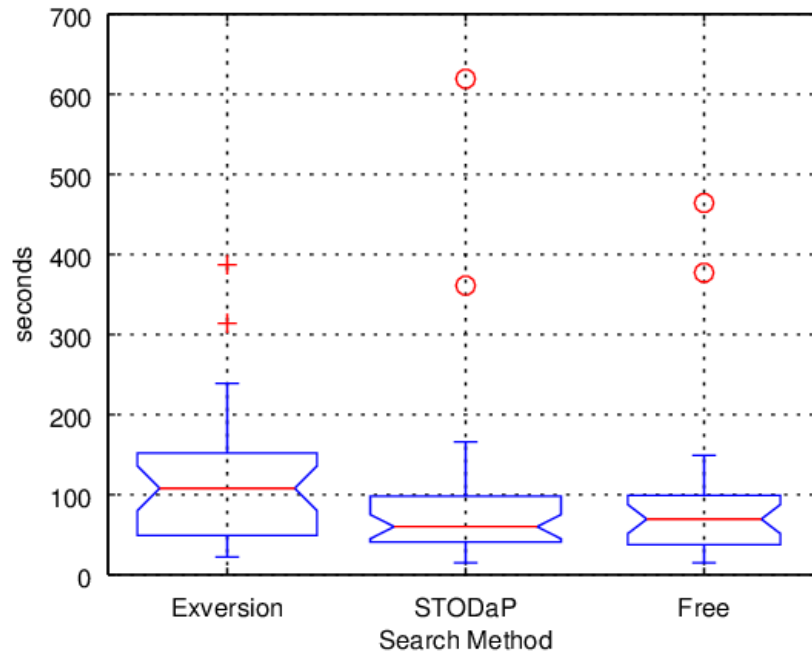


Figure 29 – Boxplot for TCT. Figure shows, for each search method: median, in red; the first and third quartiles, as the bottom and top of the box, resp.; the lowest sample within 1.5 interquartile range (IQR) of the lower quartile, as the lower whisker; the highest sample within 1.5 IQR of the upper quartile, as the upper whisker; outliers between 1.5 and 3 IQR, as + (plus); and outliers higher than 3 IQR, as o (circle).

If we look to the standard deviations on the tables, it is possible to see that they are very close to the median, and in some cases even higher. Thus, we look at the TCT for STODaP/Q2 (Table 17), one could interpret as: “When using STODaP to answer Q2, users usually take between -107 and 233 seconds to retrieve one dataset”. This obviously does not make sense. The two conclusions we can take are: (i) distribution are skewed, i.e., tails are not symmetric around the mean; and (ii) there might be outliers disturbing the standard deviation. Thus, we show in Figure 29, Figure 30 and Figure 31 the boxplot of TCT, TCT_{nb} and TCT_c , respectively.

Boxplot (BENJAMINI, 1988) pictures a more detailed view of a distribution, and is specially useful for skewed samples. Besides the median, marked with a red line in the centre of the box, Figure 29 also shows the first and third quartiles, which are the lower and upper bounds of the box. They represent, respectively, the higher among the one quarter smaller samples, and the smaller among the one quarter higher samples. Thus, half of the samples are inside the box, whose size is called interquartile range (IQR). Boxplot shows also outliers: if the distance from an outlier to the median is between 1.5 and 3 IQR, it is pointed with a circle (o); if it is higher than 3 IQR, it is marked with a plus (+).

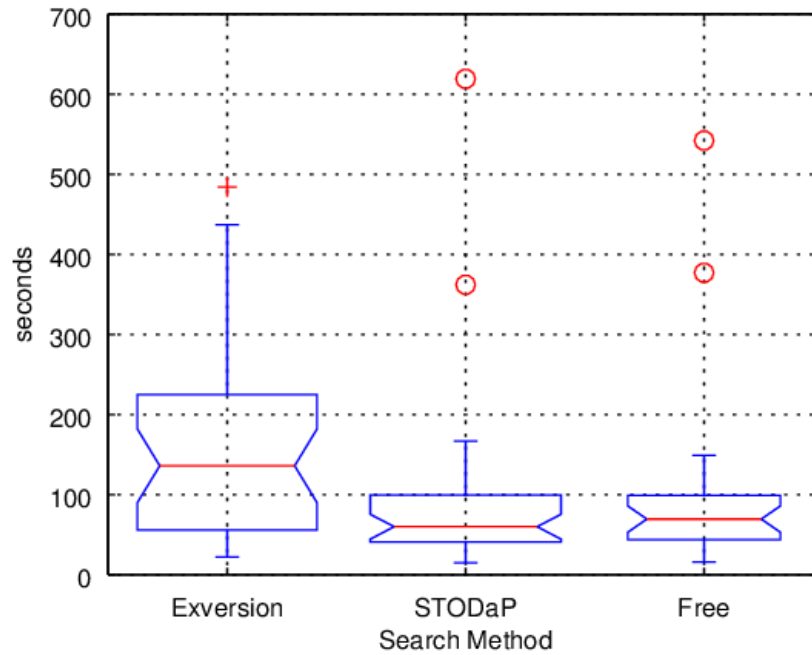


Figure 30 – Boxplot for TCT_{nb} . Elements of the plot are described in Figure 29.

Figure 29 shows that, although STODaP median is lower, IQR is almost at the same position. Free search lower bound is slightly smaller. STODaP presented higher outliers, while in Exversion outliers are less disperse. Moving to Figure 30, where blank answers are not considered, STODaP keeps the lower median, and the lower bound in this case is smaller, meaning that 50% of the samples are located in a lower region than on the free search method.

Results shown in Figure 31 strengthen this trend, and in this case STODaP distribution lies on a lower region in all aspects: median, quartiles, whiskers and outliers. In all scenarios, distribution of Exversion samples are in a higher position in relation to the other methods. On the other side, this search engine presents a better behaviour in relation to outliers.

As commented before, the better performance of STODaP in Figure 30 and Figure 31 has a direct relation to the precision of the answers. This aspect will be analysed in the following.

6.5.3 Precision Analysis

As stated before, each answer given by participants was validated against the specified task. Thus, we were able to compute the precision of each task detailed in Equação 6.2.

In order to estimate the performance of each search method in terms of precision,

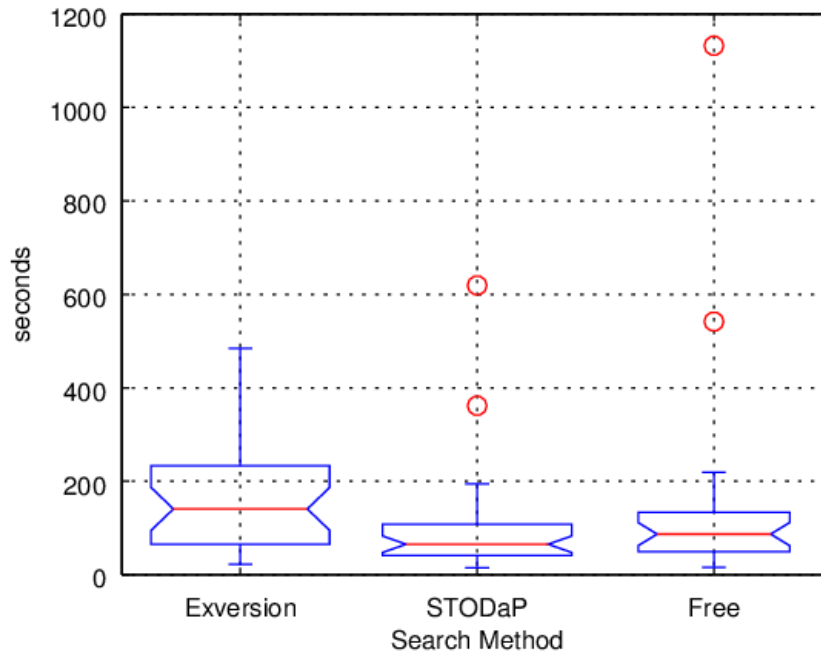


Figure 31 – Boxplot for TCT_c . Elements of the plot are described in Figure 29.

we calculate the average precision for each question, and the aggregate result for each search method. Reasons for using average and not median, as before, are the following. When measuring precision, range is limited to $33 \leq P \leq 100$ (values under 33 were discarded), and thus outliers are less troublesome than in TCT. Moreover, since some samples have many $P = 100$, median can give a less precise picture of the distribution. Figure 32 presents a bar plot of the precision averages. In order to verify the distribution behaviour, a boxplot is presented in Figure 33.

A parallel behaviour as in TCT analysis can be observed in Figure 32. STODaP exhibits a better performance on average, but is almost equivalent to free search regarding the Budget question. A 100% rate can be observed in the STODaP performance for the Procurement task. In this case, where the task required datasets in three different languages, semantic tag Public Procurement (<http://stodap.org/semantictag/4321/>) could directly connect datasets in English, Russian, Spanish, Finish, Portuguese, Danish, German and Italian using the keyword “procurement”. Thus, this task presents the better performance for STODaP, and the worse using other methods.

Figure 33 also presents some useful information. For all questions, and in the aggregated analysis, median of STODaP is 100%. The same can be stated about free search, except for Q1. IQR of STODaP is zero for Q1 and Q3 and in the aggregated result, where only 3 samples are not 100%.

Higher precision of STODaP can be explained by questions tailored to take ad-

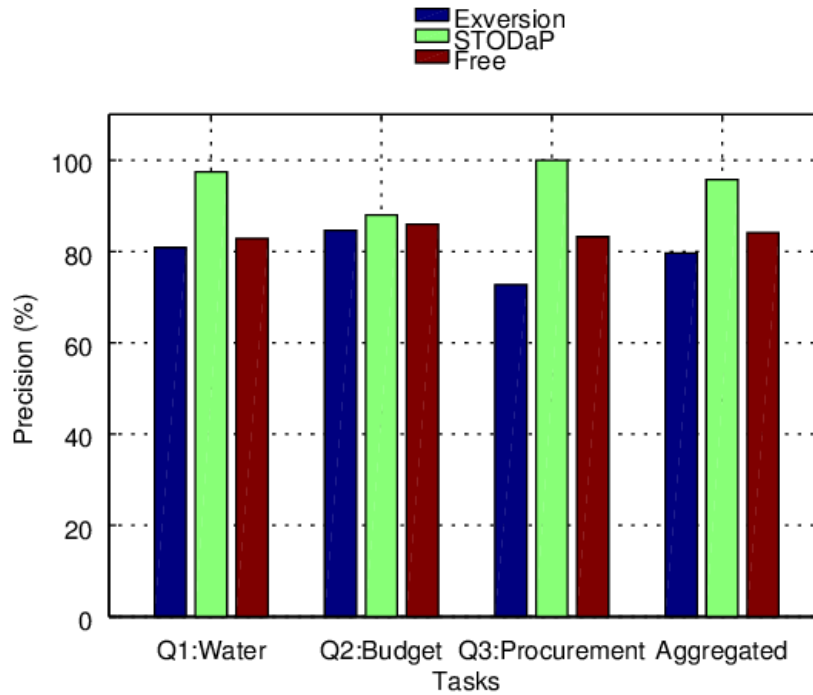


Figure 32 – Precision analysis. Bars show the average precision for participants answers, for each question/search method combination, and aggregated for each search method.

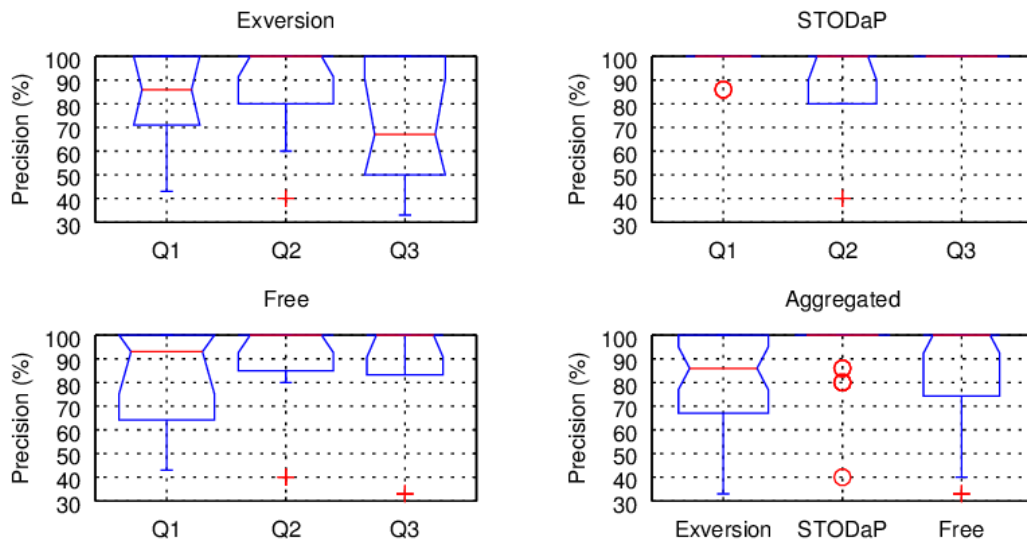


Figure 33 – Precision analysis. Figure shows a boxplot for each search method / question combinations, and aggregated for each search method.

vantage of its features. Particularly, asking for different countries and languages matches some of the STODaP facets. Results show that these features of the STODaP architecture work as expected, turning STODaP into a better search engine than the powerful generic

ones, at least for these particular cases. Further discussions are driven in the conclusions of this chapter and in the next chapter.

6.5.4 Subjective Evaluation

After completing the tasks, participants filled an evaluation questionnaire, whose answers are systematized in Table 20. The first question is aimed to gather an absolute view of participants about the system. The second one, in turn, evaluates STODaP in relation to another specialized open dataset search engine, and to generic web search engines.

Results are shown as an average of all user, but also segregated between non-experts, i.e., those who considered their open data ability being between 1 to 3, and experts, i.e., open data ability 4 or 5. Global score shows an overall satisfaction with STODaP, both absolutely and in relation to other methods. Segregated results show non-experts more satisfied than experts. This result was expected, since, as discussed before, experts normally need more specific datasets, and they know where to find it, or at least have some clues. STODaP fits more the needs of non-experts, by providing a generic starting point to find data.

Table 20 – STODaP evaluation - summary of subjective evaluation. Table shows the average results of answers to the evaluation questionnaire presented in Table 13. Answers are integers ranging from 1 (low) to 5 (high).

Question	Global Average ($n = 37$)	Non-experts ($n = 27$)	Experts ($n = 10$)
Absolute Satisfaction	4.3	4.3	4.3
Relative Satisfaction	4.2	4.3	4.0

The evaluation questionnaire included also a free comment section, in order to let participants write their impressions. Unfortunately, only 14 participants out of 37 wrote comments. Comments were categorized as:

Design/Layout: Six participants complained about the user interface, both visually and in terms of usability: *“The user interface is not intuitive nor pleasant. It should be further worked to help the visualization of results.”* was one the comments. The absence of an interface for managing filter was also noticed: *“When I select the option to filter by country, I couldn’t see any way to remove that filter or which filters are currently applied in case there are more than one.”*

Compliments: Six users wrote positive comments, highlighting the ability of the system to support proposed tasks. One example is: *“I did not manage to find any relevant datasets on river data quality after about 20 minutes of searching (using google). There were some*

available but they were for European rivers. STDOaP is very useful and simple to use. Data is very easily discovered, and results are more relevant due to the availability of filtering.”

Language: One comment was related to language barriers, specifically when search terms are not in English. The translation of semantic tags label is referred in the future works.

Missing data: Two participants noticed the absence of desired data, but one of them recognized that this could be due to a publishing failure, and not related to the STODaP system itself.

Semantic Interpreter: One subject noticed that *“some improvement in semantic interpreter is needed.”*, but unfortunately no further details were given.

6.5.5 Correlation Analysis

In order to check the correlation between characteristics of the participants, their evaluation of the tool and their performance on the tests, we calculate the correlation coefficient between every variable related to the participants, i.e.: (i) profile variables such as age, English proficiency, and internet, data and open data abilities; (ii) evaluation about the tool, i.e.: absolute and relative satisfaction; and (iii) test performance, i.e.: total tasks completion time and precision.

Table 21 shows the results. Age is measured in years. The six following columns are variables that scale from 1 to 5, where 1 refers to the worse option, and 5 to the better. Task Completion Time is the sum of time, in seconds, taken to complete all 3 tasks, and the last column is related to the percentage of correct answers. As before, for this analysis, we considered only participants with more than 33% of correct answers.

Looking to the correlation coefficients, only a few have a module higher than 0.5, and higher values do not reveal any unexpected results. It is possible to observe a positive correlation between absolute and relative satisfaction. A negative correlation between open data ability and relative satisfaction confirms the fitness of the system for non-experts in open data. Data ability is negatively correlated with task completion time, and positively with precision. English proficiency may also play a positive role regarding precision.

Although some clues can be gathered looking at Table 21, an analysis of variance (ANOVA) reveals that only a few correlations have a high confidence. The variables whose correlation confidence (p-value) is equal or than 0.05 are: age and data ability (0.001), age and relative satisfaction (0.003/0.0001), English proficiency and data ability (0.042), data and open data ability (0.029/0.02), absolute and relative satisfaction (0.002/0.003). Regarding the experiment measures – TCT and precision – no relevant correlation was found.

Table 21 – Correlation analysis of the results.

Variables	Age	Internet	English	Data	Open Data	Absolute Satisfaction	Relative Satisfaction	TCT	Precision
Age	1.0	–	0.18	0.52	0.41	-0.11	-0.42	-0.16	0.13
Internet	–	–	–	–	–	–	–	–	–
English	0.18	–	1.0	0.39	0.43	-0.22	-0.31	0.11	0.28
Data	0.52	–	0.39	1.0	0.45	-0.18	-0.28	-0.32	0.41
Open Data	0.41	–	0.43	0.45	1.0	-0.21	-0.44	-0.13	0.15
Absolute Satisfaction	-0.11	–	-0.22	-0.18	-0.21	1.0	0.56	0.15	0.14
Relative Satisfaction	-0.42	–	-0.31	-0.28	-0.44	0.56	1.0	0.15	0.08
TCT	-0.16	–	0.11	-0.32	-0.13	0.15	0.15	1.0	0.09
Precision	0.13	–	0.28	0.41	0.15	0.14	0.08	0.09	1.0

6.6 Discussion

Evaluation is a crucial part of every scientific work, since without testing, it is impossible to know if the system works as designed, and if it accomplishes the proposed goals. However, designing and implementing an information system evaluation is a very complex and challenging task. Isolating the desired variables from other environment influences is nearly impossible, mainly because of the socio-technical characteristic of information systems. And if we try to select the ideal participants to perform the ideal tasks, the evaluation itself can become too complex, and chances are high that it gets too distant of reality. One can be tempted to look for simple cause-effect explanations in complex problems, i.e., involving humans and systems. However, according to [Morin \(2011\)](#), the world is an inseparable tissue of actions, interactions, feedbacks, determination and chances, and thus it is too simplistic to think in a direct cause-effect explanation for complex systems.

Searching for open datasets is not an everyday task for the absolute majority of the population. On the other hand, experienced data scientists already know where to find their data, and even if this data is available or not. With those questions in mind, we designed this evaluation trying to balance specificity and generality in the choice of subjects. We also tried to balance the specificity of the questions, so that it would not be

too general (“Find open datasets in Brazil”) but also not too specific (“Find the open data set of 2015 budget in Rio de Janeiro”). Following this reasoning we chose our participants (Computer Science students that could work one day with open data and Semantic Web researchers and practitioners that already work with open data), and our tasks (finding datasets on specific themes according to language and geographical limitations).

As quickly discussed in Section 6.1, all those choices imply in limitations to our experiment in terms of conclusion possibilities. Some dimensions of the STODaP approach were not directly assessed, for example: ODP extensions, search by navigation (instead of keyword search), or the quality of semantic lifting. It is also worth noting that the full potential of semantic metadata was not tested. The aspect of merging several text tags into a single semantic resource⁷ was definitely responsible for the good performance of STODaP, especially when dealing with multi-country and multi-language questions. However, the extracted relations between semantic tags, such as broader, narrower and related, were not tested. Their use is more explicit in the navigation search (budget tag points to finances, as broader tag, and to budget policy, as narrower) and with ODP extensions (allowing dataset recommendation based on related tags).

6.7 Conclusions

Results show that, for the designed tasks, STODaP outperforms Exversion Data Search Engine and generic web search engines regarding time to complete tasks and precision of the answers. It can also be noticed that STODaP performance is better when less datasets are available. In cases where datasets availability is high, such as budget datasets, web search engines may find more precise data in less time.

Subjective evaluation about the system was positive, both regarding absolute usefulness, and in relation to other search methods. Participants with a lower open data experience tend to rate the system better than experts. Statistical correlation analysis confirmed that participants with a higher data manipulation ability finish their tasks faster and with a higher precision. English proficiency seems also to play a positive role regarding precision. However, no interesting correlation could be confirmed by a relevance test.

Comments written by participants point out several enhancement suggestions, e.g. regarding layout and design, semantic interpretation and multi-language enhancements. The absence of data for some topics was also noticed, recognizing this gap as a governmental transparency issue.

⁷ For example, the semantic tag *budget* - <http://stodap.org/semantictag/3643/> merges datasets of almost 40 ODPs tagged with tags like budget, haushalt, orçamento, presupuesto, talousarvio, Presupuesto, Presupuesto económico, haushaltsplan, budge, presupuestal, Budget, bilancio, Haushalt, Haushaltsplan, Orçamento or бюджет.

In the next chapter, we summarize our contributions, indicating the limits of this research and pointing out the way for continuing this work.

7 Conclusions

The challenge of turning open data more accessible is of high priority if the aim of this movement is to enhance participation and strengthen democracy, as alleged. However, as seen on this thesis, several problems still have to be overcome, and a number of perils threaten the success of open data. In this chapter, the conclusions of this thesis are derived, based on the initial hypothesis and in the contributions that were made during this work. Limitations and difficulties found on the thesis development are also discussed. Finally, we point out various directions for researchers willing to continue this work.

7.1 Contributions

As described in [Section 1.4](#), this thesis has a main objective, and also other specific ones. In this section, we first detail the main contribution, analysing the validation of our initial hypothesis and the possible generalisations of our results. The second part describes other contributions related to the specific objectives.

7.1.1 Main Contribution

In the Introduction of this work, a hypothesis was posed:

H1: *Cleaning up, reconciling and enriching metadata leads to a higher searchability of open datasets.*

The first chapters of this thesis emphasized the importance of open data in our society, and we showed by literature review ([Section 2.7](#)), participatory research ([Section 3.4](#)) and objective metrics ([Section 4.2](#)) that description of open datasets is a relevant problem. The STODaP approach proposed in [Chapter 5](#) targeted precisely the open data organisation problem, both from a local perspective (inside a single ODP) and from a global perspective (inter-ODPs). The implementation of STODaP server was evaluated in the previous chapter, and results shown that the system enabled people to find datasets more precisely and faster than using other approaches. This can be considered true specially for non-experts in data processing and open data, and regarding topics for that open datasets are not abundant.

In order to validate the hypothesis, general theoretical and practical analysis were developed, as described above. However, experiments are intrinsically specific. Thus, it is necessary to discuss if the results described in this thesis can be generalised, or if they are true only for a specific context.

As discussed in the last chapter, evaluation of STODaP server has two specific design choices that may affect the generality of the results. The first aspect was the choice of participants. Participants of the experiment can be divided in three groups: (i) first year Computer Sciences students, in Brazil; (ii) Semantic Web researchers of various nationalities; (iii) open data practitioners/activists in Brazil. Even though there are significant variations on the level of data/open data experience among these participants, all of them share a daily use of internet, an average to high English proficiency, and an average data processing ability.

The second design choice was regarding the questions that participants had to answer: (i) find 2015 open budget data from 5 countries; (ii) find open data about water quality in seven rivers outside Europe; and (iii) find open procurement datasets in 3 different languages. The choice of the topics was based on the 13 criteria used by the Open Data Index¹, in order to guarantee the relevance of the questions on the open data community.

In this context, regarding the choice of participants, it is not possible to guarantee the extension of our results to participants completely unaware of data processing, with a low English proficiency, or with a low computer/internet ability. Regarding the topics, since all Open Data Index topics have a significant amount of open datasets, it is secure to generalise our findings to other relevant open data topics. As seen in last chapter, a higher availability of datasets favours generic web search engines. No conclusions can be drawn about less popular topics, but we can speculate that STODaP would have a good performance as long as datasets exist and are adequately tagged.

Thus, about the current state of STODaP implementation, we can affirm that:

For users with at least an intermediate level of English, daily internet use, and average data experience, STODaP open data search engine delivers open datasets with a higher precision in less time than other search methods when searching for relevant open data topics.

7.1.2 Other Contributions

Chapter 3 presents some contributions on the Data Literacy field. The importance of data literacy efforts was emphasized in Section 3.2, where an analogy between traditional and data literacies was derived. We concluded showing the importance of data literacy on giving voice to marginalized people, in the same way as it was crucial to alphabetise people some decades ago. A definition of the concept of *Critical Data Literacy* was also presented,

¹ The Open Data Index compiles an open data ranking, whose criteria are: 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 of these topics, countries are evaluated and receive a score according to openness level.

emphasizing the need of a real understanding of what is behind data. In [Section 3.3](#) our proposal for working on data literacy with social movements activists was presented. A methodology for data literacy course was detailed, mixing theory, discussion and practice, in an effort to bring data literacy knowledge closer to each educands reality.

Thus, on the Data Literacy field, we bring:

- A theoretical contribution regarding the contributions of Popular Education theory to Data Literacy, and the definition of the concept of Critical Data Literacy;
- A methodological contribution, regarding the description of a Data Literacy course and associated research methodology, with an emphasis on a critical understanding and use of data by social movement activists; and
- A practical contribution, regarding the systematisation of impediments, benefits and improvements of open data according to social movement activists.

These efforts resulted in a stronger motivation for developing the STODaP approach. Observing in practice the difficulties in finding and using open data enabled this work to ground not only in problems observed in the literature, but also on the field.

[Chapter 4](#) brings also as a contribution a framework for analysing the use of tags inside ODPs and between several ODPs. Metrics developed are not innovative themselves, but their application on the ODP context bring light to the problematic use of tags.

In [Chapter 5](#), beyond the already presented contribution, there are still local strategies for cleaning up and semantically lifting tags. These contributions – the Tag Manager Plugin and the Semantic Tags Plugin – are part of the STODaP approach, but were not evaluated because their use demand modifying ODPs software, which is managed by government servants.

7.2 Limitations and Difficulties

After highlighting the contributions of this thesis, it is also necessary to expose some limitations and difficulties, and to delimit the extension of our results.

Regarding the work reported in this thesis as a whole, the main conceptual problem lies in attempting to derive a transdisciplinary approach without having the appropriate means for it. Open Data is a topic that comprehends multiple perspectives. Political sciences, regarding open data policies, legal studies, regarding access to information regulation, social sciences, regarding the effects it can have on the society, or education, regarding data literacy, are certainly a few of them. Looking at the problem from the Computer Science point of view is necessary, but the other perspectives have to be somehow included. According to the Charter of Transdisciplinarity,

The transdisciplinary vision is resolutely open insofar as it goes beyond

the field of the exact sciences and demands their dialogue and their reconciliation with the humanities and the social sciences as well as with art, literature, poetry and spiritual experience. (FREITAS; MORIN; NICOLESCU, 1994, Article 5)

The initial aim of this thesis was to achieve a balance between social and technical aspects, in the sense of connecting people's real demand to the developed approach, recognizing the complexity of the open data problem, and the need for a transdisciplinary approach. Although it was still possible to include social and educational aspects, mainly through the Data Literacy discussion, their role ended up being more motivational than really being part of the scientific development. Bringing a transdisciplinary view to the Computer Sciences field is not easy. Even when effort is put on it, as it was the case of this thesis, the absence of such an approach on the regular Computer Science education hampers the formulation of the problem and of the solution approach in a valid scientific framework. For example: how could we validate the hypothesis that, after attending to our data literacy course, an educand improved his/her capacity of critically analysing and producing data? Are there objective metrics able assess this statement? Or maybe the question should be: is this relevant to the Computer Science field or community? Unfortunately, answers to these question would require another thesis.

Besides this general remark, we can point the problem choice as another limitation of this work. By choosing the open data description problem, several others detected in [Chapter 3](#) were disregarded. It is of course impossible to deal with every problem on the open data field in a single thesis, but some of them are of crucial importance, as for example the Data Quality problem. In our evaluation procedure, participants went only until the entry point of datasets. It is possible that datasets that were marked as valid in the evaluation do not correspond to their descriptions, are outdated, or have broken links to their resources. This is a clear limitation of this work: open datasets search is based on metadata, but data quality itself is not tackled.

Another limitation of the approach is related to the subjects that participated on the evaluation. Ideally, we would expect that the same participants that attended to the Data Literacy course could evaluate the system. However, this was not possible, mainly because: (i) Most of the participants on the Data Literacy course do not speak English; (ii) the time span between the courses and the evaluation was of almost 2 years, which makes contacts more difficult. Thus, it was not possible to evaluate if the developed system attended to at least some of the requirements pointed out by participants of the courses.

Moreover, the connection between tags and semantic tags is still limited, as mentioned previously. The automatic approach used in this thesis resulted in many tags not connected to the right semantic tag, and also many wrong connections. Although several methods to tackle this issue are available in the literature, the large size of our database

requires efficient approaches that are still to be tested at STODaP.

Instability of Open Data Portals was another limiting factor. As reported in [Chapter 4](#), from 140 listed portals, only 87 could be accessed. Thus, our database could be almost twice as complete as it is now. However, due to several reasons stated before, ODP could not be accessed.

7.3 Future Work

The broadness of this work let many open paths for further research.

On the Data Literacy field, [Figure 6](#) can give relevant clues for further research. The training branch presents a great variety of topics, ranging from Basic Informatics, Mathematics and Statistics, until data journalism or open data publishing. The possibility of integrating such a wide range of topics in Data Literacy courses should be further investigated, specially considering the use of semantically agreed metadata to describe and search for open datasets.

Regarding reconciliation of tags in ODPs, a sort of crowd validation would be very useful in order to confirm automatic processing and add new information. The model proposed by [Limpens, Gandon and Buffa \(2013\)](#) could be very useful, because of its ability to deal with divergences between users. In this sense, it is necessary to call the attention of the open data community in order to further advance on collaborative strategies for enriching the semantic tags server. Another way of improving the connection between tags and groups with their semantic equivalents could be via artificial intelligence methods.

Advances on local tools for tagging assistance are crucial, since a good tagging procedure would certainly ease the work of tag reconciliation. Future research and development should include a tag suggestion approach for ODPs which takes into account the related tags at the tag server, using collective knowledge as in [Sigurbjörnsson and 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.

Another interesting research direction is detecting the emergence of schemas from the tags, as described in [Robu, Halpin and Shepherd \(2009\)](#), where tag correlation is used to create tagging vocabularies and visualizations of terms relationships. A schema emerged from ODP content and metadata could facilitate users understanding of the portal organisation.

On the semantics side, it would be of great interest to explore a richer set of relationships using pre-existing knowledge from domain ontologies. Government related ODPs share a significant set of similar domains, such as Education, Health, Transports or Budget. Using specialised ontologies of these domains could certainly improve the

qualification of relationships.

The STODaP implementation also needs extensions on its development in order to be really useful. A priority list of enhancements could start by:

- *Internationalization.* At the moment, semantic tags and groups labels are only in English. This means that users searching for keywords in English are able to find datasets in other languages via semantic tags, but the contrary is not true. The Gemet Thesaurus offers several translations of the terms, which could be used for this task.
- *User Interface Design.* Several participants commented on the evaluation that interface design should be improved. This is a priority because interface problems can hide system functionalities, and thus result in a worse performance than the system could offer.
- *Inclusion of other open data systems.* The list of ODPs indexed by the STODaP server came from the CKAN Census. The first problem is that this Census is not frequently updated, since many CKAN ODPs listed there do not exist more, and other not present in the Census were found. Secondly, a quantity of ODPs are developed using Socrata, another open data management system. Although being proprietary, Socrata also offers APIs for external users to consume data. Developing a connection to pull Socrata metadata could significantly improve the STODaP base.

* * *

For STODaP to realize its full potential, ODP administrators and users should be involved and (meta)data literacy needs to be improved. Far from being a technical problem, open data will not advance if politicians and public servants in charge of transparency policies do not give special attention to data literacy. Including critical data literacy on scholar curriculum could consistently enhance the level of data skills on the society.

Although open data is currently a trendy word, its “openness” is still limited by many factors, including political and technological ones. With this work, we hope to have contributed to the open data field, maybe not so much with the resulting tool, but hopefully with a socio-technical view where the user side can have at least the same importance as the publisher side.

References

- ADAMS, T.; STRECK, D. R. Educação Popular e novas tecnologias. *Educação*, Porto Alegre, v. 33, n. 2, p. 119–127, 2010. Cited 2 times on pages 47 and 51.
- 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 79.
- 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 44.
- 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 4 times on pages 20, 77, 78, and 79.
- 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 78.
- 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 66.
- 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 36 and 41.
- 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. Available from Internet: <<http://dx.doi.org/10.5281/zenodo.18592>>. Cited 2 times on pages 28 and 43.
- 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. Available from Internet: <<http://link.springer.com/10.1007/978-3-319-09846-3>>. Cited on page 68.
- BARATO, J. N. *Codification/decodification: an experimental investigation on the adult education theory of Paulo Freire*. Tese (Thesis) — San Diego State University, 1984. Available from Internet: <<https://jarbas.wordpress.com/048-codificacaodecodificacao-em-paulo-freire/>>. Cited on page 50.
- 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 30.

- BATES, J. The strategic importance of information policy for the contemporary neoliberal state: the case of Open Government Data in the United Kingdom. *Government Information Quarterly*, v. 31, n. 3, p. 388–395, 2014. Cited on page 19.
- 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 36.
- BENJAMINI, Y. Opening the Box of a Boxplot. *The American Statistician*, [American Statistical Association, Taylor & Francis, Ltd.], v. 42, n. 4, p. 257–262, 1988. ISSN 00031305. Available from Internet: <<http://www.jstor.org/stable/2685133>>. Cited on page 128.
- BERNERS-LEE, T. Linked Data - Design Issues. *W3C Website*, 2006. Available from Internet: <<http://www.w3.org/DesignIssues/LinkedData.html>>. Cited 2 times on pages 20 and 41.
- BERNERS-LEE, T. *5 Stars Open Data*. 2010. Available from Internet: <<http://5stardata.info/>>. Cited 3 times on pages 30, 38, and 101.
- 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 42.
- BHARGAVA, R. *Towards a Concept of "Popular Data"*. 2013. Available from Internet: <<https://datatherapy.org/2013/11/18/towards-a-concept-of-popular-data/>>. Cited on page 47.
- 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 20 and 45.
- BHATTACHARYA, J. *How Google Processes Queries in a Semantic Web Environment*. 2014. Available from Internet: <<https://ahrefs.com/blog/google-processes-queries-semantic-web-environment/>>. Cited on page 118.
- BIRD, S.; LOPER, E.; KLEIN, E. *Natural Language Processing with Python*. [S.l.]: O'Reilly Media Inc., 2009. Cited 2 times on pages 94 and 104.
- BRANDÃO, C. R. *O que é o método Paulo Freire*. São Paulo: Brasiliense, 1985. Cited on page 50.
- 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 36.
- 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 45.
- 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 79.

- 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 80.
- 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 2 times on pages 115 and 121.
- CHIGNARD, S. *A Brief History of Open Data*. 2013. Available from Internet: <<http://www.paristechreview.com/2013/03/29/brief-history-open-data/>>. Cited 2 times on pages 27 and 45.
- 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 79.
- COLPAERT, P. et al. The 5 stars of open data portals. In: *7th MeTTeG*. [S.l.: s.n.], 2013. p. 61–67. Cited 2 times on pages 81 and 91.
- 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 76.
- CORAZZA, S. M. *Tema Gerador: concepção e prática*. Ijuí: Editora Unijuí, 2003. Cited on page 50.
- 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 on page 20.
- Data Revolution Group. *A World That Counts - Mobilising the Data Revolution for Sustainable Development*. [S.l.], 2014. 32 p. Available from Internet: <<http://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf>>. Cited 3 times on pages 19, 40, and 44.
- DAVIES, T. *Open data, democracy and public sector reform*. 1–47 p. Tese (Dissertation) — University of Oxford, 2010. Cited 2 times on pages 38 and 39.
- 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 21 and 38.
- DAVIES, T.; BAWA, Z. A. The Promises and Perils of Open Government Data (OGD). *Journal Of Community Informatics*, v. 8, n. 2, p. 1–6, 2012. Cited on page 19.
- DAVIES, T.; SHARIF, R. M.; ALONSO, J. M. *Open Data Barometer - Global Report - 2nd Edition*. The World Wide Web Foundation, 2015. 1–62 p. Available from Internet: <<http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf>>. Cited 2 times on pages 32 and 40.
- 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 43.

- 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 80.
- EAVES, D. *The Three Laws of Open Government Data*. 2009. Available from Internet: <<http://eaves.ca/2009/09/30/three-law-of-open-government-data/>>. Cited on page 30.
- 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. Available from Internet: <<http://doi.acm.org/10.1145/2506182.2506196>>. Cited on page 80.
- EXVERSION. *Building a Search Engine for Data*. 2015. Available from Internet: <<https://exversiondata.wordpress.com/2015/04/21/building-a-search-engine-for-data/>>. Cited on page 117.
- FALS-BORDA, O.; RAHMAN, M. A. *Action and knowledge: breaking the monopoly with participatory action-research*. [S.l.]: Appex Press, 1991. Cited on page 44.
- FELLBAUM, C. *WordNet: An Electronic Lexical Database*. 1998. 423 p. Cited on page 85.
- 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 48.
- 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 47.
- FIORETTI, M. *A proposal to promote Open Data from and for the schools*. 2011. Available from Internet: <<http://mfioretti.com/2011/10/warsaw-open-data-and-education/>>. Cited on page 46.
- FREIRE, P. *Educação e Mudança*. 12. ed. Rio de Janeiro: Paz e Terra, 1979. Cited on page 50.
- FREIRE, P. *Pedagogia do Oprimido*. 11. ed. [S.l.]: Editora Paz e Terra, 1987. Cited on page 48.
- FREIRE, P. *Pedagogy of the Oppressed*. 30. ed. New York: Continuum, 2005. ISBN 8521902433. Available from Internet: <<http://www.infed.org/thinkers/et-freir.htm>>. Cited on page 50.
- FREITAS, L. de; MORIN, E.; NICOLESCU, B. *Charter of transdisciplinarity*. Portugal, 1994. 1–5 p. Cited on page 140.
- GADOTTI, M. *Paulo Freire: Uma Biobibliografia*. São Paulo: Cortez Editora/Instituto Paulo Freire, 1996. Cited on page 49.
- 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 51.

GRANICKAS, K. *Understanding the impact of releasing and re-using open*. [S.l.], 2013. 1–29 p. Cited on page 36.

GREY, J.; BOUNEGRU, L.; CHAMBERS, L. *Data Journalism Handbook*. [S.l.]: OKFN, 2012. Cited on page 45.

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. Available from Internet: <<http://tomgruber.org/writing/ontology-of-folksonomy.htm>>. Cited 3 times on pages 73, 93, and 100.

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 4 times on pages 19, 22, 33, and 40.

GURSTEIN, M. B. Why I'm Giving Up on the Digital Divide. *Journal Of Community Informatics*, v. 11, n. 1, 2015. Available from Internet: <<http://ci-journal.net/index.php/ciej/article/view/1210/1139>>. Cited 2 times on pages 45 and 52.

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. Available from Internet: <<http://portal.acm.org/citation.cfm?id=1242602>>. Cited on page 73.

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. Available from Internet: <<http://bioinformatics.oxfordjournals.org/content/30/5/740.abstract>>. Cited on page 79.

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. Available from Internet: <<http://dx.doi.org/10.2200/S00639ED1V01Y201504HLT027>>. Cited 2 times on pages 79 and 83.

HERN, A. *New York taxi details can be extracted from anonymised data*. 2014. Available from Internet: <<http://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn>>. Cited on page 37.

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 4 times on pages 19, 26, 45, and 46.

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 50.

JETZEK, T.; AVITAL, M.; BJØRN-ANDERSEN, N. Generating Value from Open Government Data. *ICIS 2013 Proceedings*, p. 1–20, 2013. Available from Internet: <<http://aisel.aisnet.org/icis2013/proceedings/GeneralISTopics/5>>. Cited on page 36.

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. Available from Internet: <<http://dcpapers.dublincore.org/ojs/pubs/article/view/925>>. Cited on page 74.

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 74.

KNERR, T. *Tagging ontology-towards a common ontology for folksonomies*. [S.l.], 2006. 3–8 p. Available from Internet: <<https://tagont.googlecode.com/files/TagOntPaper.pdf>>. Cited 2 times on pages 73 and 74.

KRÖTZSCH, M. et al. Semantic wikipedia. *Journal of Web Semantics*, December 2007. Available from Internet: <http://www.aifb.uni-karlsruhe.de/Publikationen/showPublikation_english?publ_id=1551>. Cited on page 102.

LANIADO, D.; MIKA, P. Making sense of Twitter. In: *ISWC*. [S.l.: s.n.], 2010. Cited on page 76.

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. Available from Internet: <<http://ieeexplore.ieee.org/xpls/abs/all.jsp?arnumber=6406217>>. Cited on page 78.

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 5 times on pages 20, 72, 79, 80, and 141.

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>. Available from Internet: <<http://dl.acm.org/citation.cfm?id=2063531>>. Cited on page 75.

MANYIKA, J. et al. *Open Data: Unlocking Innovation and Performance with Liquid Information*. [S.l.], 2013. 24 p. Cited 2 times on pages 19 and 36.

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 72, 73, 75, and 78.

MELO, G. D. Lexvo . org : Language-Related Information for the Linguistic Linked Data Cloud. *Semantic Web Journal*, v. 7, p. 1–5, 2015. Cited on page 85.

MIKA, P. Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics*, v. 5, n. 1, p. 5–15, 2005. ISSN 15708268. Cited on page 73.

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 33.

MORIN, E. *Introdução ao Pensamento Complexo*. 4a. ed. [S.l.]: Editora Sulina, 2011. Cited on page 134.

MURRAY-RUST, P. Open Data in Science. *Serials Review*, v. 34, p. 52–64, 2008. ISSN 00987913. Cited on page 26.

- NAVARRO, G. A guided tour to approximate string matching. *ACM Computing Surveys*, v. 33, n. 1, p. 31–88, 2001. ISSN 03600300. Available from Internet: <<http://portal.acm.org/citation.cfm?doid=375360.375365>>. Cited on page 77.
- NEWMAN, R. *Tag ontology design*. 2005. Available from Internet: <<http://www.holygoat.co.uk/projects/tags/>>. Cited on page 73.
- 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 50.
- OBAMA, B. *Memorandum for the Heads of Executive Departments and Agencies*. The White House, 2009. Available from Internet: <<https://www.whitehouse.gov/the-press-office/TransparencyandOpenGove>>. Cited on page 27.
- 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 76.
- Open Knowledge Foundation. *Open Data Handbook*. [S.l.]: OKFN, 2015. Cited on page 26.
- OPENSPEEDING. *Budget Data Package*. [S.l.], 2014. Available from Internet: <<https://github.com/openspending/budget-data-package>>. Cited 2 times on pages 33 and 35.
- 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 3 times on pages 19, 37, and 38.
- 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 2 times on pages 74 and 85.
- 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 76.
- RENZIO, P. D.; WEHNER, J. *The Impacts Openness: A Review of the Evidence*. [S.l.], 2015. 35 p. Cited on page 36.
- 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. Available from Internet: <<http://eprints.soton.ac.uk/268192/>>. Cited on page 141.
- 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 2 times on pages 19 and 41.

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 52.

SCHIELD, M. Information Literacy, Statistical Literacy and Data Literacy. *IASSIST Quarterly Summer/Fall*, v. 28, n. 2/3, p. 6–11, 2004. Available from Internet: <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.144.6309>>. Cited 2 times on pages 21 and 46.

School of Data. *School of Data Report 2014*. 2014. Available from Internet: <<http://2014.schoolofdata.org/>>. Cited on page 46.

SCHUGURENSKY, D. *Paulo Freire*. London: Bloomsbury Publishing, 2014. Cited on page 49.

SCHULER, D.; NAMIOKA, A. *Participatory Design: Principles and Practices*. [S.l.]: Lawrence Erlbaum Associates, 1993. Cited on page 44.

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. Available from Internet: <<http://dl.acm.org/citation.cfm?id=1367542>>. Cited on page 141.

SMITH, B. Ontology. *Blackwell Guide to the Philosophy of Computing and Information*, n. 1964, p. 155–166, 2003. ISSN 1943-4723. Cited on page 42.

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. Available from Internet: <<http://www.springerlink.com/content/413285327hj53234/>>. Cited 4 times on pages 20, 77, 78, and 79.

TAUBERER, J. *Open Government Data: The Book (2nd Edition)*. [S.l.]: Author's Edition, 2014. Cited 3 times on pages 27, 29, and 43.

The World Bank. *Open data for economic growth*. [S.l.], 2014. 1–20 p. Cited 2 times on pages 27 and 36.

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 79.

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 4 times on pages 24, 33, 102, and 127.

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. Available from Internet: <<http://arxiv.org/abs/1510.04501>>. Cited 2 times on pages 24 and 87.

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. Available from Internet: <<http://ci-journal.net/index.php/ciej/article/view/1220/1165>>. Cited 8 times on pages 24, 48, 55, 58, 64, 157, 158, and 159.

- 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 3 times on pages 24, 48, and 54.
- 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 76, 82, 84, and 91.
- 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. Available from Internet: <<http://dl.acm.org/citation.cfm?id=2506182.2506201>>. Cited on page 33.
- 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 46.
- 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 3 times on pages 20, 77, and 78.
- 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 on page 116.
- 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 33.
- VLEUGELS, R. Overview of all FOI laws. *Fringe Special*, p. 1–28, 2012. Cited 2 times on pages 18 and 32.
- 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 80.
- 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 116.
- 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. Available from Internet: <<http://oro.open.ac.uk/43855/1/webSci-CR.pdf>>. Cited on page 46.
- WORTHY, B. *David Cameron's Transparency Revolution? The Impact of Open Data in the UK*. London, 2013. Cited on page 33.
- 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. Cited on page 73.

- 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 on page 115.
- YANNOUKAKOU, A.; ARAKA, I. Access to Government Information: Right to Information and Open Government Data Synergy. *Procedia - Social and Behavioral Sciences*, Elsevier B.V., v. 147, p. 332–340, 2014. ISSN 18770428. Available from Internet: <<http://www.sciencedirect.com/science/article/pii/S187704281404018X>>. Cited on page 18.
- ZUIDERWIJK, A.; JANSSEN, M. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, Elsevier B.V., v. 31, n. 1, p. 17–29, 2014. ISSN 0740624X. Available from Internet: <<http://dx.doi.org/10.1016/j.giq.2013.04.003>>. Cited on page 19.
- 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. Available from Internet: <<http://doi.acm.org/10.1145/2612733.2612761>>. Cited 4 times on pages 33, 37, 38, and 40.
- ZUIDERWIJK, A. et al. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, v. 10, n. 2, p. 156–172, 2012. Cited 7 times on pages 19, 20, 21, 37, 39, 40, and 41.

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. 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. 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. ; KIRSCH, R. . Contributions of Paulo Freire for a Critical Data Literacy: a Popular Education Approach. , to appear in Journal of Community Informatics.
- TYGEL, A. F. ; CAMPOS, M. L. M. ; ALVEAR, C. A. S. . Teaching Open Data for Social Movements: a Research Strategy. Journal of Community Informatics, v. 11, p. 1, 2015.
- TYGEL, A. F. ; GONÇALVES, L. G. ; SANTOS, M. ; MARQUES, G. ; CAMPOS, M. L. M. . 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.

A.4 Special Issue Co-editor

- FRANK, M. ; WALKER, J. ; ATTARD, J. ; TYGEL, A. F. . Journal of Community Informatics – Data Literacy Special Issue. GURSTEIN, M. and VILLANUEVA-MANSILLA, E. (Eds.).

APPENDIX B – Results of Open Data Research

Table 22 – 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, Campos and Alvear (2015)

Table 23 – 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 organisation is confusing
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, Campos and Alvear \(2015\)](#)

Table 24 – 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, Campos and Alvear \(2015\)](#)