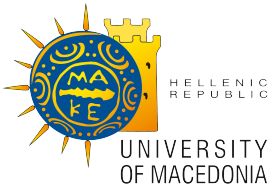
University of Macedonia



Doctoral Thesis

Linked Open Government Data Analytics

Author:
Evangelos Kalampokis

Supervisor:
Prof. Konstantinos Tarabanis

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

Interdepartmental Programme of Postgraduate Studies in Information Systems

June 24, 2016

Πανεπιστημιο Μακεδονιας

Διδακτορικη Διατριβη

Αναλυτική Συνδεδεμένων Ανοικτών Κυβερνητικών Δεδομένων

Συγγραφέας:
Ευάγγελος Καλαμποκης

Επιβλέπων:
Κωνσταντίνος Ταραμπανης

Διατριβή υποβληθείσα προς εκπλήρωση των απαραίτητων προϋποθέσεων για την απόκτηση Διδακτορικού Διπλώματος

στο

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών στα Πληροφοριακά Συστήματα

June 24, 2016

«Σκοπὸς τῆς ζωῆς μας δὲν εἶναι ἡ χαμέρπεια. Ὑπάρχουν ἀπειράκις ὡραιότερα πράγματα καὶ ἀπ' αὐτήν τὴν ἀγαλματώδη παρουσία τοῦ περασμένου ἔπους. Σκοπὸς τῆς ζωῆς μας εἶναι ἡ ἀγάπη. Σκοπὸς τῆς ζωῆς μας εἶναι ἡ ἀτελεύτητη μᾶζα μας. Σκοπὸς τῆς ζωῆς μας εἶναι ἡ λυσιτελής παραδοχή τῆς ζωῆς μας καὶ τῆς κάθε μας εὐχῆς ἐν παντί τόπῳ εἰς πᾶσαν στιγμήν εἰς κάθε ἔνθερμον ἀναμόχλευσιν τῶν ὑπαρχόντων. Σκοπὸς τῆς ζωῆς μας εἶναι τὸ σεσημασμένον δέρας τῆς ὑπάρξεώς μας.»

University of Macedonia

Abstract

Interdepartmental Programme of Postgraduate Studies in Information Systems

Doctor of Philosophy

Linked Open Government Data Analytics

by Evangelos KALAMPOKIS

Public sector produces, collects, maintains, and disseminates a wealth of data. It is widely recognised the potential of exploiting these *government data* to boost among others economic activity, innovation and public administration transparency.

In 2009, responding to the call of Sir Tim Berners-Lee, inventor of the world wide web, governments worldwide started to massively make data available online in open licenses and technical formats that facilitate reuse. They launched *Open Government Data (OGD)* portals that operate as single points of access for government data.

The focus of this thesis is the OGD movement and its contribution in realising the potential of government data. Towards this end, we study OGD in a holistic approach by taking into account the viewpoints of both providers and consumers.

Initiatives that provide OGD are part of public sector and as such they inherit deficiencies coming from the decentralised organisational structure of public sector, which comprises multiple administrative levels and functional areas. Moreover, the technological formats and the structure of data that are provided through OGD portals affect data exploitation. *Linked Data* has been early proposed as the most advanced technological paradigm for opening up data because it facilitates data integration across the Web. Moreover, aggregated statistics (e.g. economic and social indicators) structured as multi-dimensional data cubes constitute a major part of OGD.

On the other hand, consumers perceive OGD as a small fraction of massive amounts of data that are daily produced and made available on the Web from various sources such as social media, research institutions, and news media. These data are provided in different technological formats and some times with diverse access constraints. In this emerging reality, the *integration* of OGD with other Web data is of vital importance for addressing the needs of consumers. Moreover, we consider that OGD exploitation has to capitalise on the paradigm of *data analytics*, which has already enabled organisations

to successfully exploit their own data in various problem areas such as business intelligence.

Within this problem formulation in this thesis we explore (a) *provision*, (b) *integration*, and (c) *exploitation in data analytics* of OGD and we propose specific solutions, including conceptual models, architectures, and software tools, that contribute towards realising the full potential of government data. The proposed solutions are evaluated in scenarios that involve real-world datasets from OGD portals, social media, clinical trials, etc. Since OGD movement emerged only recently we ground our analysis in traditional conceptual models of electronic government.

The contribution of this thesis can be summarised as follows:

- Provision
  - An OGD classification scheme that provides an understanding of the domain.
  - An OGD stage model that can be used as a roadmap for future endeavours.
  - A process model that describe the lifecycle of multi-dimensional OGD.
- Integration
  - Architectures and implementations for integrating OGD and social media data on the Linked Data Web.
  - A theoretical framework for integrating multi-dimensional OGD.
  - An analysis of the challenges for integrating multi-dimensional OGD on the Linked Data Web.
- Exploitation in Data Analytics
  - A set of software tools that enable performing online analytical processing (OLAP) analytics on top of multiple datasets across the Linked Data Web.
  - A study of performing exploratory analytics on top of integrated data for elections understanding.
  - A process model that enables exploiting social media data in predictive analytics. The model was used to evaluate the predictive power of social media and to design a case for predicting winner of 2010 UK elections using integrated Twitter and linked open data.
- Access
  - An access control framework that enables combining open and private data on the Linked Data Web.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Επιτομή

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών στα Πληροφοριακά Συστήματα

Διδακτορικό Δίπλωμα

Αναλυτική Συνδεδεμένων Ανοικτών Κυβερνητικών Δεδομένων

Ευάγγελος Καλαμποκης

Ο δημόσιος τομέας παράγει, συλλέγει, συντηρεί και διανέμει πληθώρα δεδομένων. Είναι κοινά αποδεκτή η δυναμική της αξιοποίησης των κυβερνητικών δεδομένων για την ενίσχυση, μεταξύ άλλων, της οικονομικής δραστηριότητας, της καινοτομίας, και της διαφάνειας στην δημόσια διοίκηση.

Το 2009, ανταποκρινόμενες στην πρόσκληση του Sir Tim Berners-Lee, εφευρέτη του παγκόσμιου ιστού, οι κυβερνήσεις σε όλο τον κόσμο άρχισαν να διαθέτουν μαζικά τα δεδομένα τους χρησιμοποιώντας ανοικτές άδειες και τεχνικές μορφοποιήσεις που διευκολύνουν την επαναχρησιμοποίηση. Ίδρυσαν πύλες *Ανοικτών Κυβερνητικών Δεδομένων (ΑΚΔ)* οι οποίες λειτουργούν ως μοναδικό σημείο πρόσβασης για κυβερνητικά δεδομένα.

Το επίκεντρο αυτής της διατριβής είναι το κίνημα των ΑΚΔ και η συμβολή του στην υλοποίηση της δυναμικής των κυβερνητικών δεδομένων. Προς το σκοπό αυτό, μελετούμε τα ΑΚΔ με μία ολιστική προσέγγιση, λαμβάνοντας υπόψη την οπτική τόσο των παρόχων όσο και των καταναλωτών.

Οι πρωτοβουλίες που παρέχουν ΑΚΔ αποτελούν μέρος του δημόσιου τομέα και συνεπώς κληρονομούν ελλείψεις που προέρχονται από την αποκεντρωμένη οργανωτική δομή του δημοσίου, η οποία περιλαμβάνει πολλαπλά επίπεδα διοίκησης και λειτουργικές περιοχές. Επιπλέον, οι τεχνολογικές μορφοποιήσεις και η δομή των δεδομένων που παρέχονται μέσω των διαδικτυακών πυλών ΑΚΔ επηρεάζουν την αξιοποίηση των δεδομένων. Τα συνδεδεμένα δεδομένα (linked data) έχουν από νωρίς προταθεί ως το πιο προηγμένο τεχνολογικό παράδειγμα για το «άνοιγμα» των δεδομένων στον Ιστό. Επίσης, συγκεντρωτικά στατιστικά (π.χ. οικονομικοί και κοινωνικοί δείκτες) τα οποία δομούνται ως πολυ-διάστατοι κύβοι αποτελούν ένα σημαντικό μέρος των ΑΚΔ.

Από την άλλη πλευρά, οι καταναλωτές αντιλαμβάνονται τα ΑΚΔ ως ένα μικρό κλάσμα από τις τεράστιες ποσότητες δεδομένων που παράγονται και διατίθενται καθημερινά στον ιστό από διάφορες πηγές όπως τα μέσα κοινωνικής δικτύωσης, τα ερευνητικά ιδρύματα, και τα

μέσα ενημέρωσης. Αυτά τα δεδομένα παρέχονται με διαφορετικές τεχνολογικές μορφοποιήσεις και κάποιες φορές με ποικίλους περιορισμούς πρόσβασης. Σε αυτή τη νέα πραγματικότητα, η σύνδεση των ΑΚΔ με άλλα δεδομένα του Ιστού είναι απαραίτητη για την ικανοποίηση των αναγκών των καταναλωτών. Επίσης, θεωρούμε ότι η αξιοποίηση των ΑΚΔ θα πρέπει να κεφαλαιοποίηση το παράδειγμα της *αναλυτικής δεδομένων (data analytics)*, το οποίο έχει ήδη επιτρέψει σε οργανισμούς να αξιοποιήσουν τα δικά τους δεδομένα σε ποικίλες περιοχές όπως στην επιχειρηματική ευφυΐα.

Μέσα σε αυτήν την διαμόρφωση του προβλήματος στην παρούσα διατριβή διερευνούμε (α) την παροχή, (β) την ολοκλήρωση, και (γ) την αξιοποίηση με αναλυτική δεδομένων των ΑΚΔ και προτείνουμε συγκεκριμένες λύσεις που περιλαμβάνουν θεωρητικά μοντέλα, αρχιτεκτονικές, και εργαλεία λογισμικού, τα οποία συμβάλουν προς την πραγματοποίηση της πλήρης προοπτικής των κυβερνητικών δεδομένων. Οι προτεινόμενες λύσεις αξιολογούνται σε σενάρια που περιλαμβάνουν σύνολα δεδομένα από ΑΚΔ πύλες, μέσα κοινωνικής δικτύωσης, ερευνητικά πειράματα, κλπ. Καθώς το κίνημα των ΑΚΔ αναδύθηκε μόλις πρόσφατα βασίζουμε την ανάλυση μας σε παραδοσιακά θεωρητικά μοντέλα της ηλεκτρονικής διακυβέρνησης.

Acknowledgements

First and foremost I want to thank my supervisor Prof. Konstantinos Tarabanis for inviting me to become a member of his team and giving me the opportunity to get involved in research. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience rich and productive.

Profound gratitude goes to Ass. Prof. Efthimios Tambouris for his constant support and coaching especially during my first steps.

I would also like to thank all my colleagues in the Information Systems Laboratory at the University of Macedonia and in the Digital Enterprise Research Institute (DERI). Special mention goes to the following:

- Dr. Michael Hausenblas (formerly at DERI) for his contagious and motivational joy and enthusiasm for Linked Data.
- Ms. Areti Karamanou for getting her hands dirty with hundreds of thousands of tweets.
- Ms. Eleni Kamateri that contacted employees at clinical research organisations (CING, CHUV and ZEINCRO) and identified requirements regarding clinical research data access and exploitation in the course of the EU funded Linked2Safety research project.
- Mr. Dimitrios Zeginis that contributed to the development of the OLAP Browser and proof-of-concept platform of LiMDAC.
- Mr. Paul Hermans that organised the evaluation of the OLAP Browser from employees at the Flemish Government in the course of the EU funded OpenCube research project.

Contents

List of Figures

List of Tables

Abbreviations

CSO	**C**entral **S**tatistics **O**ffice
CSV	**C**omma **S**eparated **V**alues
DCLG	**D**epartment for **C**ommunities and **L**ocal **G**overnment
EU	**E**uropean **U**nion
FOAF	**F**riend **O**f a **F**riend
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
JSON	**J**ava**S**cript **O**bject **N**otation
IWB	**I**nformation **W**ork**b**ench
LiMDAC	**Li**nked **M**edical **D**ata **A**ccess **C**ontrol
LIMO	**LI**nked Statistical **MO**del Vocabulary
LOD	**L**inking **O**pen **D**ata
MA	**M**oving **A**verage
NER	**N**amed **E**ntity **R**ecognition
OGD	**O**pen **G**overnemnt **D**ata
OKFN	**O**pen **K**nowledge **F**oundation
OLAP	**O**n**l**ine **A**nalytical **P**rocessing
PNR	**P**ositive **N**egative **R**atio
PSI	**P**ublic **S**ector **I**nformation
QB	RDF data **Q**u**b**e
RDBMS	**R**elational **D**atabase **M**anagement **S**system
RDF	**R**esource **D**escription **F**ramework
ROR	**R**egistry **O**f **R**esources
SKOS	**S**imple **K**nowledge **O**rganization **S**ystem
SM	**S**ocial **M**edia
SPARQL	**S**PARQL **P**rotocol **A**nd **R**DF **Q**uery **L**anguage

UK	**U**nited **K**ingdom
URI	**U**niform **R**esource **I**dentifier
URL	**U**niform **R**esource **L**ocator
US	**U**nited **S**tates
W3C	**W**orld **W**ide **W**eb **Consortium**
XML	E**X**tensible **M**arkup **L**anguage

στὴν μνήμη τοῦ ἀδελφοῦ μου Παναγιώτη

Chapter 1

Introduction

1.1 Motivation

Public sector produces, collects, maintains and disseminates a wealth of information. Governments all over the world realise that *"within the exercise of its public tasks, the public sector collects, processes and disseminates huge quantities of information"* (European Commission, 2003). Examples include maps and satellite images, legislation and case-law, statistics and company, population and patent registers (European Commission, 2009). In Europe, *"public bodies are by far the largest producers of information"* (European Commission, 2000).

The availability of this information (government data onwards) in easily accessible digital format makes it possible to re-use it and combine it with other digital content to create new added-value services and products. Examples include navigation services, real-time traffic information, weather forecasts e.g. sent directly to mobile phones and credit rating services (European Commission, 2009). It is widely recognised that such data-based, added value services and products increase government transparency, improve public administration's function, contribute to economic growth and provide social value to citizens (Burdon, 2009; Dekkers et al., 2006). They generate new businesses and jobs and give consumers more choice and more value for money (European Commission, 2009).

The value of the government data market in the European Union (EU) is estimated having a mean value around 27 billion Euros (Dekkers et al., 2006). More than a decade

ago the European Commission recognised the potential of exploiting this information to boost economic activity and job creation (European Commission, 1998, 2001). At the political level, the European Parliament and the Council have launched a Directive on the re-use of government data (European Commission, 2003). Therefore, government data constitutes a valuable asset for both society and economy and as a result governments have a mandate to enable and facilitate data consumption and exploitation by both citizens and businesses.

Nevertheless, problems on government data re-use such as lack of information on available data (European Commission, 2009) or the need to bring some order to the mass of data produced still exist. A recent evaluation of the European Directive underpins a number of barriers towards the full exploitation of government data (European Commission, 2009). Things seems better in USA where re-use is strongly encouraged (European Commission, 2009) however even there the potential of government data has not been fully exploited.

This situation seems to change in the last years, where a large number of governments worldwide started to massively make data available on the Web. This *Open Government Data* (OGD) movement follows the *Open Data* philosophy suggesting making data freely available to everyone, without limiting restrictions. One of the main tenets of OGD is that government provides data and then private parties built added value products and services that provide interactive access for the public (Robinson et al., 2008).

1.2 Problem Definition

However, a simplistic view is often adopted with regards to OGD, which automatically correlates the publicising of data with use and benefits and thus the potential of OGD has been unrealised to a large extend (Janssen et al., 2012).

The difficulty in exploiting OGD seems surprising if we consider the huge importance data have in modern societies. Indeed, during the last years, businesses, academia and government employ various data analytics methods on their own data with great success. For example, business intelligence methods are employed by enterprises to help them survive in the global economy (Jourdan et al., 2008). In addition, evidence based

policy-making relies on data analytics to assist policy makers in producing better policies (Sanderson, 2002).

This difficulty could be explained by a number of barriers (legal, political, social, organisational, and technical) that hamper the interaction between public administration and society (citizens and enterprises). The former publishes OGD in an ad-hoc manner based on existing processes, according to their mandate, and often under unclear licenses. They also design and deliver services in a top-down manner. On the other hand, society has needs and data-driven public services, not raw data, can address these needs (Peristeras and Tarabanis, 2000).

Moreover, government data is produced, collected, stored and disseminated by public agencies. Each agency manages data according to its mandate. The issues related to re-use of government data would be much easier resolved from an organisational perspective, if public agencies (a) were totally independent from each other and (b) were managing different data from those managed by other agencies. However, in the public sector none of these two conditions is true.

On the contrary, agencies formulate hierarchical structures that contain a number of administrative levels. Thus, agencies have in their responsibility and sometimes control other agencies, i.e. those belonging to a lower administration level. In addition, the public sector is organised in functional areas, such as education, health etc. This decentralised organisational structure of the public sector suggests that in certain cases public agencies in different administration levels and different functional areas produce, maintain and possibly disseminate similar data i.e. data about similar real-world objects or problems. This situation results in a number of challenges regarding data quality. In particular, it is possible that the disseminated data is incomplete, controversial and/or obsolete.

At the same time, the Web is moving from a model of connected documents to a model based on the connections between real-world objects and data describing these objects [1]. In this context, a number of Web sites and platforms opened up recently their data. Examples include Facebook's Graph API , Twitter's RESTful API , the semantically enabled Google's Rich Snippets and also the Linking Open Data project , which realised the provision of Linked Data from a number of Web sources such as Wikipedia. Linked

[1] http://dig.csail.mit.edu/breadcrumbs/node/215

Data aims to extend the Web with a data commons by creating typed links between data from different sources (Bizer et al., 2009a).

Government data are part of this ongoing evolution of the Web and thus it should be combined and integrated with other open data on the Web in order to allow for added value services. To this end, both governments and private sector are expected to develop the necessary technological infrastructure and establish the appropriate organisational processes. Governments could be involved and play an important role in this process because they own the data and thus can understand it better than third parties.

In this emerging reality it is evident that there is a lack of a proper understanding and roadmap guidelines to set clear objectives for next steps and benchmarks and measure progress in the OGD movement.

This is particularly true, as the various stage models developed during the last decade for measuring the progress of eGovernment development do not seem appropriate for OGD. Indeed, these models often consider online information provision as the lowest stage in eGovernment development while the higher stages aim at enabling online transaction and providing sophisticated online services through governments transformation (Layne and Lee, 2001; Lee, 2010; Siau and Long, 2005). Apparently, the existing eGovernment stage models are not capable of describing the increasing OGD movement.

1.3 Scope of Research

The scope of this thesis is to study the recently emerged Open Government Data (OGD) movement. More precisely, our study is focused on exploring whether and how the OGD movement can realise the potential of government data. Towards this end, we study OGD in a holistic approach by taking into account the viewpoints of both providers and consumers of OGD.

Regarding OGD provision, we study OGD portals as part of public sector and as such we consider that they inherit deficiencies coming from the decentralised organisational structure of public sector, which comprises multiple administrative levels and functional areas. As a result, we focus on organisational challenges related to OGD provision. From a technological point of view we focus on *Linked Data* as a paradigm that facilitates data

integration on the Web. Linked data has been early proposed as the most advanced and promising way for opening up OGD (e.g. see (Berners-Lee, 2009)). As a result, we adopt linked data as the technological paradigm that underpins our implementations and guides our technological analyses. Moreover, we emphasise on aggregated statistics (e.g. economic and social indicators) structured as multi-dimensional data cubes because these data constitute a significant part of the available open data provided through OGD portals. It is indicative that the vast majority of the datasets published on the open data portal of the European Commission are of statistical nature.

Regarding OGD exploitation, on this thesis we focus on creating added-value through data integration and analytics. We study OGD integration both inside and outside public sector. In the fist case, we study the integration of data from public agencies in different administrative levels or functional areas that, however, refer to same real world problem. In the second case, we consider data from other sources on the Web. In particular, we exploit data from (a) social media data, (b) news media, (c) clinical researches, and (d) DBpedia, i.e. the linked open data version of Wikipedia. We also employ data analytics methods in order to create value from integrated data. The term "data analytics" is often used broadly to cover any data-driven decision making. In this thesis we focus on analysing data using typical methods that are used for years in business intelligence and analytics. We exploit online analytical processing (OLAP) as well as well established statistical analysis and data mining techniques for association analysis, classification and regression analysis, and predictive modelling. As a result, this thesis is not about developing new statistical analysis or data mining techniques.

We should finally note that, the OGD movement aims to unlock public information to enable re-using it and combining it with other digital content to create new added-value services and products. However, there are a number of important challenges for realising this aim. These include legal issues, such as those relevant to data privacy and protection, cultural issues, e.g. related to politicians and public servants, and socio-technical issues related to organisational and technological challenges. In this thesis, we concentrate on the latter issues particularly those related to enable re-using government data including combining it with other open data on the Web.

1.4 Related Work

As already suggested, Open Government Data (OGD) initiatives emerged only recently. In many cases however OGD is part of electronic government (eGovernment) and more specifically online one-stop government and governmental portals. Therefore, in this section we first present work related to eGovernment emphasising online one-stop government data portals and relevant classification schemes as then proceed with work related to OGD.

1.4.1 eGovernment stage models

During the last decade, a number of models and schemes have been suggested by international organisations, consulting firms and researchers in order to provide a roadmap for eGovernment development and to enable evaluation of relevant initiatives. The European Union proposed a five-stage maturity model in order to enable benchmark and rate "governments service delivery processes" (Lorincz et al., 2009). The stages included in the model, which are described based on maturity and sophistication, are the following: information, one-way interaction, two-way interaction, transaction and finally targetisation. Layne and Lee (2001) in order to describe different stages of eGovernment development introduced a "stage of growth model for fully functional eGovernment". This model comprises four stages, namely cataloguing, transaction, vertical integration and horizontal integration. These stages are explained in terms of organisational and technological complexity as well as different levels of integration. Deloitte Research described the stages that a government will pass as electronic service delivery evolves (Breen, 2000). The aim of this model was to identify the key issues governments need to resolve to make this moving successful. The proposed model includes six stages, namely information publishing/dissemination, official two-way transactions, multi-purpose portals, portal personalisation, clustering of common services, full integration and enterprise transformation. Deloitte Research described the model using two axes: the eminence of web-based applications and the degree of enterprise transformation. eGovernment transformation was described by West (2004) using four stages, namely the bill board stage, the partial service delivery stage, the portal stage, including fully executable and integrated service delivery, and interactive democracy including public outreach and accountability enhancing features. West's aim was to provide a tool to researchers to

determine an agency's progress based on how far along they are at incorporating various web site features. To this end, he studied more than 1800 government websites in the United States and carried out a survey involving chief information officers in different state and federal agencies. Based on Layne and Lee's model, Andersen and Henriksen (2006) proposed a progressive growth model for eGovernment. Here, the key dimensions are the degree of activity-centric websites and processing of end-users information and service requests. The first phase of the model is cultivation, which shelters horizontal and vertical integration within government, limited use of front-end systems for customer services and adoption and use of Intranet within government. The next phase is extension that involves extensive use of intranet and adoption of personalised Web user interface for customer processes. Phase three is maturity where the organisation matures and abandons the use of the intranet, have transparent processes, and offers personalised Web interface for processing of customer requests. The last phase is revolution characterised by data mobility across organisations, application mobility across vendors, and ownership to data transferred to customers. In this phase, the employees actions can be traced through the Internet and there is information available online about progress in, for example, case handling. Khalil et al. (2002) suggested a model to divide the process of eGovernment implementation into three independent phases. These phases do not need one phase be completed before another can begin. The first one is publish, i.e. using ICT to expand access to government information, the second is interact, i.e. broadening civic participation in government, and the last one is transact, i.e. making government services available online.

In addition, work has been also carried out aiming to compare and synthesise eGovernment models. For example, Siau and Long (2005) developed a five-stage model to synthesise eGovernment stage models of that time so that to create a common frame of reference for researchers and practitioners in the area. This model is described in terms of time, complexity and integration as well as benefits and costs. More specifically, according to this model time spending, system complexity, integration, benefits and costs all increase with the advancement of eGovernment. The model proposed consists of five stages, namely web presence, interaction, transaction, transformation and eDemocracy. Finally, Lee (2010) also compared existing eGovernment development models in order to identify a common frame of reference across different stage model. This framework comprises two dimensions, namely citizen/service perspective and operation/technology

TABLE 1.1: Review of eGovernment stage models (adopted from Kalampokis et al. (2011c))

Model	*Dimensions*	*Stages*
Andersen and Henriksen (2006)	Degree of activity-centric websites and processing of the end-users information and service requests	Cultivation, extension, maturity and revolution.
Khalil et al. (2002)	n/a	Publish, interact and transact.
Breen (2000)	Eminence of web-based applications and the degree of enterprise transformation	Information publishing, two-way transactions, multi-purpose portals, portal personalisation, clustering of common services, full integration and enterprise transformation.
Lorincz et al. (2009)	Maturity and sophistication	Information, one-way interaction, two-way interaction, transaction and targetisation.
Layne and Lee (2001)	Organisation/technological complexity and different levels of integration	Cataloguing, transaction, vertical integration and horizontal integration.
Lee (2010)	Citizen/service and operation/technology	Presenting, assimilating, reforming, morphing and eGovernance.
Siau and Long (2005)	Time/complexity/integration and benefits/costs	Web presence, interaction, transaction, transformation and eDemocracy.
West (2004)	n/a	Bill board, partial service delivery, portal stage (with fully executable and integrated service delivery) and interactive democracy.

perspective and five stages, namely presenting, assimilating, reforming, morphing and eGovernance.

Table 1.1 summarises the review of the existing eGovernment stage models. As regards the dimensions utilised for describing the model the majority of the works include dimensions related to socio-technical issues such as technological complexity, organisational complexity, enterprise transformation etc. As regards the description of data provision the majority of these models consider it as the first stage in eGovernment development. The table reveals that the various stage models developed during the last decade for measuring the progress of eGovernment development do not seem appropriate

for OGD. Indeed, these models often consider online information provision as the lowest stage in eGovernment development while the higher stages aim at enabling online transaction and providing sophisticated online services through governments transformation. Apparently, the existing eGovernment stage models are not capable of describing the increasing OGD movement.

1.4.2 Open government data models

Moreover, during the last couple of years and after the start of this thesis a few conceptual process models describing open data and linked data have been also proposed.

In particular, open data processes specify the steps that governments should follow to set their data free for others to reuse (Curtin, 2010; Janssen and Zuiderwijk, 2012; van Veenstra and van den Broek, 2015). For example, the process introduced by Janssen and Zuiderwijk (2012) involves five steps, namely creating data, opening data, finding open data, using open data, and discussing and providing feedback on open data.

Moreover, linked data processes describe the steps that one have to follow in order to publish and consume linked data (Auer et al., 2013; Ding et al., 2012; Hyland and Wood, 2011; Marx et al., 2013; Villazón-Terrazas et al., 2011). For example, Auer et al. (2013) present a process comprising eight steps: (i) transform data to RDF, which includes the extraction of data from sources (structured or unstructured) and its mapping to an RDF data model, (ii) store and index data efficiently and using appropriate mechanisms, (iii) manual revise, extend and create new structured information according to the initial data, (iv) establish links to different sources that regard the same entities but are published by different data publishers (v) enrich data with high-level structures so as to be more efficiently aggregated and queried (vi) assess data quality using data quality metrics available for structured information such as accuracy of facts and completeness, (vii) repair data so as to encounter data quality problems identified in the previous step, and (viii) search, browse and explore the data in a fast and user friendly manner.

These generic models are relevant to the work that we have conducted in our thesis although they have been introduced after we start working on the area. These models can be though as complementary to the models that we introduce in these thesis as they describe data analysis processes in different granularities. The structure of our thesis

and our overall approach is based on such a lifecycle with three phases i.e. provision, integration, and exploitation of OGD. Here we should note that during our work on this thesis we have identified a gap regarding the conceptualisation of a lifecycle for multi-dimensional data that are modelled as linked data.

1.5 Outline of the Thesis

The remainder of this thesis is structured as follows:

Chapter 2 sets the background for the thesis by describing the different types of data that are currently available on the Web. In particular, it defines the concept of *Open Data* (section 2.2), while it elaborates on *Open Government Data (OGD)* as a rich source of open data (section 2.3), and *Linked Data* as a promising technical paradigm for opening up data on the Web (section 2.4). It also differentiates two categories of open data that present distinct requirements, namely (a) record-level and (b) aggregated multi-dimensional data (section 2.5). Finally, it describes two other sources of data on the Web, namely *Social Media* and *Research*, which produce data that do not fully comply with the open data definition but can be potentially combined with open data in order to increase latter's value (section 2.6).

Chapter 3 presents the OGD framework that provides a holistic understanding (both technical and organisational) of OGD. The framework comprises two conceptual models, namely a classification scheme for OGD (section 3.2) and a stage model for OGD (section 3.3). The classification scheme is based on relevant literature and enables identifying, analysing and classifying OGD initiatives. We believe that having a classification scheme allows a deeper understanding of initiatives and therefore the domain as a whole. It comprises two main dimensions. The first dimension refers to the technological approach followed for making data available on the Web, while the second refers to the organisational approach followed for providing OGD. The stage model aims at (a) providing a roadmap for open government data re-use and (b) enabling evaluation of relevant initiatives' sophistication. The proposed model has two main dimensions, namely organisational & technological complexity and added value for data consumers.

Chapter 4 explores OGD integration both inside and outside public sector. Towards this end, we take into account structural and technical formats of OGD. Regarding structural formation of data we differentiate between record level data and aggregated multi-dimensional data. From a technological point of view we employ linked data paradigm in our analysis. As a results, this chapter explores technical and organisational challenges of two cases of OGD integration: (a) combining record level data on the linked data Web (section 4.2), and (b) combining aggregated multi-dimensional data on the linked data Web (section 4.3). In the first case, an architecture that enables OGD integration around real-world things is presented along with a proof-of-concept implementation that provides an integrated view of OGD from different public authorities in different administration levels (sub-section 4.2.1). Moreover, the case of combining OGD with social media data is explored and a technical architecture based on linked data along with a case study are presented (sub-section 4.2.2). In the second case, in which integration of aggregated multi-dimensional data cubes is explored, a theoretical framework that formally describes OGD integration is presented. This theoretical framework comprises a conceptual process model, a theoretical algebra, and technical recommendation for applying the model and the algebra in the linked data Web.

Chapter 5 explores how integrated open data can be exploited in data analytics in order to create added value. This chapter introduces the concept of linked OGD analytics (section 5.2) and elaborates on two types of analytics, namely OLAP analytics (section 5.3) and predictive analytics (section 5.4). In the first case, the potential of performing OLAP operations on top of integrated views of multiple data sets on the linked data Web is explored. Towards this end, we describe an innovative linked data OLAP browser that we have developed and we demonstrate its functionalities by combining and analysing OGD published by two governments in Europe, namely the Flemish and the Scottish governments. At the last section of this chapter we focus on predictive analytics as a promising way of exploiting data on the Web. In this case, we give particular emphasis on Social Media data because they incorporates personal opinions, thoughts, and behaviours making them a vital component of the Web and fertile ground for a variety of business and research endeavours. In particular, we initially explore the predictive power of SM and define a process model for performing predictive analytics on the Web. Based on these, we design a case study that aims at predicting the winning party of UK elections 2010 by exploiting combined Twitter and linked open data.

Chapter 6 explores how access constraints affect the exploitation of data on the Web. This chapter focuses on aggregated multi-dimensional data produced by clinical researches and proposes a solution that enables combining open and private data across distributed sources on the linked data Web with diverse access constraints. In particular, it presents a framework comprising (a) three linked data models, namely a metadata model for describing aggregated medical data, a user profile model for describing medical data consumers, and an access policy model, and (b) an architecture that exploits and orchestrates the three models to enable controlling access to clinical research data. From a technological perspective, the framework is validated using a proof-of-concept platform that is developed for that purpose.

Finally, Chapter 7 recalls the main findings of the research and discusses open questions emerging from this work.

1.6 Contribution

The contribution of the thesis can be organised in four categories and summarised as follows:

- Provision of open data (Chapter 3), which includes:
  - An early understanding of the Open Government Data (OGD) landscape around the globe (Kalampokis et al., 2011b).
  - A stage model for OGD that describes how value can be created by combining OGD with other data on the Web such as Social Media data (Kalampokis et al., 2011c).
  - A process model for creating, integrating, and exploiting OGD (Tambouris et al., 2015).
- Integration of open data (Chapter 4), which includes:
  - Architectures and implementations that describe the integration of record level OGD (Kalampokis et al., 2011b, 2013d) as well as OGD and SM data (Kalampokis et al., 2011a) in order to create value.

  - A theoretical framework for integrating multi-dimensional open data (Kalampokis et al., 2016c).
  - An in-depth analysis of the challenges that hamper integration of OGD structured as multi-dimensional cubes in the Linked Data Web (Kalampokis et al., 2015).

- Exploitation of integrated open data (Chapter 5), which includes:
  - A set of software tools that enable performing OLAP analytics on top of integrated views of multiple datasets across the Linked Data Web (Kalampokis et al., 2014, 2016c,d).
  - A study of performing exploratory analytics on top of integrated data for elections understanding (Kalampokis et al., 2013c).
  - A process model that enables exploiting open data from SM in predictive analytics. The process model was used to enable an in-depth understanding of the predictive power of SM (Kalampokis et al., 2013b). Moreover, it has been used to design a case study for predicting 2010 UK elections using integrated Twitter and linked open data (Kalampokis et al., 2016a).

- Access constraints (Chapter 6), which includes:
  - An access control framework that enables combining open and private data on the Linked Data Web. The framework, which comprises models and architecture, was implemented in the case of medical data integration (Kamateri et al., 2014).

Parts of this thesis have been published or have been submitted for publication in international refereed journals and conferences. In Appendix A the references are documented along with citations to these publications from journals that are indexed by Thomson Reuters' *Web of Science*.

Chapter 2

Web Data

2.1 Introduction

This chapter sets the background for the thesis by describing the different types of data that are currently available on the Web. In particular:

- It defines the concept of *Open Data* (section 2.2) and elaborates on (a) *Open Government Data* as a rich source of open data (section 2.3) and (b) *Linked Data* as a promising technical paradigm for opening up data on the Web (section 2.4).
- It differentiates two categories of open data that present distinct requirements, namely (a) record-level and (b) aggregated multi-dimensional data (section 2.5).
- It describes two sources of data, namely *Social Media* and *Research*, which produce data that do not fully comply with the open data definition but can be potentially combined with open data in order to increase latter's value (section 2.6).

2.2 Open Data

2.2.1 Definition of Open Data

The term *"Open Data"* springs from some of the same roots as *"Open Source"* or *"Open Access"*. Although *"Open"* in software normally means libre (i.e. free in the sense of

having no restrictions), there is an increasing movement towards using *"Open Access"* to mean gratis (i.e. free in the sense of costing no money) and not libre. The GNU project suggests that Open Source (or Free) software is a matter of liberty, not price, and means that *"the users have the freedom to run, copy, distribute, study, change and improve the software"*[1].

The European Commission defines open data as referring to the idea that certain data should be freely available for re-use (European Commission, 2011). This represents the use of the data for purposes foreseen or not foreseen by the original creator.

The World Bank categorises the conditions that open data have to satisfy in two broad categories:

- *Technically open*: available in a machine-readable standard format, which means it can be retrieved and meaningfully processed by a computer application
- *Legally open*: explicitly licensed in a way that permits commercial and non-commercial use and re-use without restrictions.

McKinsey Global Institute suggests that open data share the following characteristics (Manyika et al., 2013):

- *Accessibility*: A wide range of users is permitted to access the data.
- *Machine readability*: The data can be processed automatically.
- *Cost*: Data can be accessed free or at negligible cost.
- *Rights*: Limitations on the use, transformation, and distribution of data are minimal.

Recently, the Open Knowledge Foundation (OKF) introduced the *"Open"* definition[2] that applies to a piece of knowledge (or work) including (a) content (e.g. music, films, books), (b) data (e.g. scientific, historical, geographic or otherwise), and (c) government or other administrative information (software is excluded in this definition). According to the definition the following conditions should be met:

[1] https://www.gnu.org/philosophy/free-sw.html
[2] http://opendefinition.org/od/

- Access: the work should be available as a whole and at no more than a reasonable reproduction cost. It must also be available in a convenient and modifiable form.
- Redistribution: the license shall not restrict any party from selling or giving away the work either on its own or as part of a package made from works from many different sources.
- Reuse: the license must allow for modifications and derivative works and must allow them to be reattributed under the terms of the original work.
- Technological restriction: The work must be provided in such a form that there are no technological obstacles to the performance of the above activities. This can be achieved by the provision of the work in an open data format, i.e. one whose specification is publicly and freely available and which places no restrictions monetary or otherwise upon its use.
- No Discrimination: the license must not discriminate against any person, group or specific field or endeavour.

For the purposes of this thesis, open data is as defined by the OKF[3]: *"Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike"*.

2.2.2 Value of Open Data

A study conducted by the McKinsey Global Institute estimated the global annual economic potential value of Open Data to $3 trillion (Manyika et al., 2013). This figure estimates the economic value without taking into account the potential societal benefits. In addition, this study splits this potential economic value at seven areas of the global economy, namely education, transportation, consumer products, electricity, oil and gas, health care and consumer finance. The order that these areas are mentioned denotes their value from the highest to the lowest.

The study conducted by the McKinsey Global Institute (Manyika et al., 2013) splits the open data potential economic value at seven areas of the global economy, namely education, transportation, consumer products, electricity, oil and gas, health care and

[3] http://opendefinition.org

consumer finance. The order that these areas are mentioned denotes their value from the highest to the lowest.

2.2.3 Categories of Open Data

OKFN also categorises open data in eight categories, namely:

- Geodata: the data that is used to make maps. Initiatives in this category include OpenStreetMap[4].
- Culture: data about cultural works and artefacts.
- Science: data that is produced as part of scientific research.
- Financial: data such as government accounts and information on financial markets
- Statistics: data produced by statistical offices such as the census and key socioeconomic indicators.
- Weather: the many types of information used to understand and predict the weather and climate.
- Environment: information related to the natural environment such as presence and level of pollutants, the quality of rivers and seas.
- Transport: data such as timetables, routes, on-time statistics.

While the World Bank categorises open data into the following topics:

- *Agriculture & Rural Development* includes measures of agricultural inputs, outputs, and productivity compiled by the UN's Food and Agriculture Organisation.
- *Aid Effectiveness* covers indicators related to aid received as well as progress in reducing poverty and improving education, health, and other measures of human welfare.
- *Climate Change* includes data that cover climate systems, exposure to climate impacts, resilience, greenhouse gas emissions, and energy use.

[4] http://openstreetmap.org

- *Economy & Growth* covers measures of economic growth, such as gross domestic product (GDP) and gross national income (GNI) and indicators such as capital stock, employment, investment, savings, consumption, government spending, imports, and exports.
- *Education* includes data on education inputs, participation, efficiency, and outcomes.
- *Energy & Mining* includes data on energy production, use, dependency, and efficiency.
- *Environment* related data covers forests, biodiversity, emissions, and pollution.
- *External Debt* refers to data that takes a closer look at the external debt of high-income countries and emerging markets.
- *Financial Sector* includes data regarding indicators such as size and liquidity of stock markets, the accessibility, stability, and efficiency of financial systems.
- *Gender* refers to data on gender at the country level that covers demography, education, health, labor force and employment, and political participation.
- *Health* refers to data that cover health systems, disease prevention, reproductive health, nutrition, and population dynamics.
- *Infrastructure* include data regarding investments in water, sanitation, energy, housing, transport, and information and communication technologies.
- *Poverty* covers poverty and inequality measures.
- *Private Sector* covers data from the Doing Business Indicators or Enterprise Surveys.
- *Science & Technology* refers to data about research and development, scientific and technical journal articles, high-technology exports, royalty and license fees, and patents and trademarks.
- *Social Development* covers child labor, gender issues, refugees, and asylum seekers.
- *Social Protection & Labor* The supply of labor available in an economy includes people who are employed, those who are unemployed but seeking work, and first-time job-seekers. Not everyone who works is included: unpaid workers, family

workers, and students are often omitted, while some countries do not count members of the armed forces.

- *Trade* is a key means to fight poverty, improve developing country access to markets, and support a rules based, predictable trading system.
- *Urban Development* may include data on urbanisation, traffic and congestion, and air pollution

2.3 Open Government Data

Public sector bodies (or institutions) collect, produce, hold and disseminate information while accomplishing their public tasks. Examples range from geographical information, tourist information, statistics, weather data, data from publicly funded research projects, public sector budgeting, and performance levels to digitized books from libraries and all kinds of data about policies and inspection.

This information is often called Public Sector Information (PSI) (e.g. Deloitte, 2013; European Commission, 1998; Vickery, 2011), Public Data (e.g. European Commission, 2011), Public Information (e.g. of Fair Trading, 2006), and Government Data (e.g. Cameron, 2010). In this thesis we use the term *government data* to refer to this information.

As a result, we use onwards use the term *Open Government Data (OGD)* to refer to government data in which the open data definition applies.

Often the term *Open Data* is used to describe this information. For example, Janssen et al. define open data as *"non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage or distribution"*.

Finally, we should note that the European Commission differentiates between government data of administrative and non-administrative nature (European Commission, 1998). The first category includes information about administrative procedures while the second other information that can be extremely important of the decisions of firms (e.g. statistical, financial and geographic information).

2.3.1 OGD Value

The importance and potential value of government data was described a long ago by the EU (European Commission, 1998). In addition the benefits of opening up government data has been recently studied by academia (e.g. Janssen et al., 2012).

According to the European Commission (European Commission, 2011) OGD can be exploited to provide value towards two main directions: (a) developing added value services and products, (b) addressing societal challenges, and (c) accelerating scientific progress. In addition, based on the finding of an empirical study (Janssen et al., 2012), which included a group session and interviews, the benefits of OGD can be categorised in three clusters: (a) political and social, (b) economic, and (c) operational and technical. In the same context, Deloitte (2013) defined different types of value generated by OGD according to the beneficiary i.e. data provider, data user and the wider society.

According to a recent study that was conducted by Deloitte (2013) in the United Kingdom the value of OGD to consumers, businesses and the public sector in 2011–2012 was approximately £1.8 billion. Adding the social value this figure gives an aggregate estimate of between £6.2 billion and £7.2 billions.

The wider societal value can be related to (European Commission, 1998; Janssen et al., 2012):

- More transparency
- Democratic accountability
- More participation and self-empowerment of citizens. For example, a graduate student created a new algorithm for public school assignment in Boston, using information released by the city on the quality and location of schools Shi (2013).
- Creation of trust in government
- New government services for citizens
- Creation of new insights in the public sector
- Improvement of policy-making processes

- increasing democratic participation i.e. allow citizens to make more informed choices and interact with policymakers to challenge their assumptions and improve the policy making process
- promoting greater accountability
- essential for the mobility of both workers and categories like students and retired people e.g. within the EU. A better knowledge of opportunities, circumstances and procedures in countries throughout Europe can help them to make more informed choices about mobility.

OGD can be used by businesses, individuals and the public sector to:

- stimulate innovation and develop new products and services.
- hold public service providers to account, promote democratic engagement and foster greater transparency and better policy making.
- reduce barriers to entry into markets and address information asymmetries; The absence of accessible information may create a competitive disadvantage for the foreign firms compared with local firms that can draw upon their own experience on the local situation European Commission (1998).
- generate network effects that drive disruptive change by connecting increasing numbers of consumers and businesses.
- public procurement. Access to information on the local situation is necessary to make the rules work efficiently and to optimise fair chances for all firms involved. European Commission (1998).
- better accessibility of information on the state of the art of research, could decrease the amount spent on research that has been done before European Commission (1998).

2.3.2 Technologies

OGD initiatives emerged only recently and as a result, there is a lack of academic studies to analyse them based on classification schemes. There is however an increasing number

of practical guidelines suggested by various stakeholders. We present in this sub-section two sets of guidelines which in our view constitute the most influential approaches.

The World Wide Web Consortium (W3C) e-Government Interest Group suggest three steps for public administrations to open and share their data[5]:

- firstly, publish data in raw form by means of files in well-known and non-proprietary formats such as CSV and XML,
- next, create online catalogues of the raw data, and
- finally, make the data machine-readable.

Sir Tim Berners-Lee invited governments to publish data according to linked data principles (Berners-Lee, 2009). He further proposed a five-star maturity model as follows (Berners-Lee, 2010):

- 1 star: publishing data on the Web even in proprietary and desktop-centric formats under an open license. In this case data might be locked-up in a a PDF document.
- 2 stars: publishing data in a structured way (machine-readable formats) even if proprietary software is required to access data (e.g. Excel documents).
- 3 stars: publishing data in machine-readable and non-proprietary formats using open standards, e.g., CSV.
- 4 stars: publishing data using linked data principles (see section 2.4 for a more detailed description on linked data).
- 5 stars: linking the available data to other data to provide context.

2.3.3 Challenges

A number of challenges have been identified in the literature regarding the wide exploitation of OGD:

[5] http://www.w3.org/TR/gov-data/

- Lack of feedback loops regarding the integration of input (data-driven content, data-driven services) into governmental decisions or operations (Evans and Campos, 2013). Data publishers have no policies or procedures in place to follow up on the use of their data, identify interesting applications of their data and evaluate the insights potentially relevant for improving their policies in any particular area, or simply interact and participate in the network of users (activists, researchers, journalists etc.) interested in data-driven knowledge.

- Expertise and resources were found to be the most important driving factors of participation of (business) users in the context of open data. In order to attract a more varied participant base data publishers need to provide incentives to make open data interesting, relevant, and promising to individuals and organisations.

- That public organisations had not accounted to the needs and requirements of prospective users. Open data published is not always interesting, useful, or of sufficient quality for developers; a more reliable and sustainable data supply and maintenance is needed in order for the entrepreneurs to transform data to services.

- The so-called 'data divide' (Gurstein, 2011) reveals the tension between providing comprehensive datasets and meeting the needs of users who have varying levels of knowledge background and skills.

- Ensure that the open data process is aligned with real life societal problems and is actually delivering on its promise.

- There are limitations regarding the skills of potential citizen users of data such as language barrier, insufficient information literacy, and lack of domain knowledge (Martin et al., 2013).

- There are a variety of contexts which play a part in the open data process including legal (Kulk and Van Loenen, 2012) and cultural, and the variety of data content and types. Different types of data, with different content, may need a different legal, cultural, or technical treatment.

- The publication and use processes of open data are complex (Helbig et al., 2012; Zuiderwijk and Janssen, 2014) and it is not easy to predict how users will use open data, when they will use it, and how it will be used in the future (Meijer et al., 2014). For example, government agencies may avoid publishing open data as a

risk aversion strategy. The risk aversion is driven by their uncertainty in how the data will ultimately be used or combined with other data.

- Institutional barriers (e.g. No uniform policy for publicising data), task complexity barriers (e.g. Data formats and data sets are too complex to handle and use easily), use and participation barriers (e.g. No statistical knowledge or understanding of the potential and limitations of statistics), legislation barriers (e.g. No license for using data), information quality barriers (e.g. Unclear value), technical barriers (e.g. Absence of standards) (Janssen et al., 2012)

- Legislative barriers (e.g. privacy, standards, licensing), Economic barriers (e.g. value, cost, incentives), Access barriers (e.g. culture, public safety, accessibility, planning, capabilities) (Deloitte, 2013):

2.3.4 OGD initiatives

In Table 2.1, the names and the URLs of 24 initiatives are presented along with the responsible authorities. This list includes initiatives commenced by public authorities and initiatives aiming at creating a portal acting as a single point of access for data originated by disparate public authorities. We should note that this is a list of OGD initiatives that was published in 2011 by Kalampokis et al. (2011b). For a comprehensive and up-to-date list of OGD portals one can consider DataPortals.org[6]. It is indicative of the wide adoption of the OGD movement the fact that from 2011 to 2015 the number of OGD portals across the globe was increased from 24 to 519.

Table 2.1 suggests that initiatives have been established in federal (e.g., the Federal US government and the Australian government), national (e.g., the UK government), regional (e.g., the State of California and the Victorian government), and local (e.g., the City of Vancouver and the District of Columbia) level.

[6] http://dataportals.org

TABLE 2.1: List of OGD Initiatives (adopted from Kalampokis et al. (2011b))

Name	**URL**	**Responsible authority**
Catalogo de Datos de Asturias	http://risp.asturias.es	The Principality of Asturias, Spain
City of Edmonton Open Data catalogue	http://data.edmonton.ca	City of Edmonton, Canada
Data.australia.gov.au	http://data.australia.gov.au	The Australian Government
Data.ca.gov	http://data.ca.gov	State of California
Data.gov	http://data.gov	The Federal US Government
Data.gov.uk	http://data.gov.uk	The UK Government
Data.govt.nz	http://data.govt.nz	New Zealand government
Data.nsw	http://data.nsw.gov.au	New South Wales Government
Data.seattle.gov	http://data.seattle.gov	The Seattle City Government
Data.vic.gov.au	http://data.vic.gov.au	The Victorian Government
DataSF	http://www.datasf.org	The City of San Francisco
Dati.piemonte.it	http://www.dati.piemonte.it	Region of Piedmont, Italy
Dc.gov data catalogue	http://data.octo.dc.gov	District of Columbia
Lichfield Open Data	http://lichfielddc.gov.uk/data	Lichfield District, UK
London datastore	http://data.london.gov.uk	Greater London Authority
Maine.gov DataShare	http://www.maine.gov/data	State of Maine
Mass.gov/data	http://mass.gov/data	The Commonwealth of Massachusetts

Continued on next page

Table 2.1 – continued from previous page

Name	URL	Responsible authority
NYC Data Mine	http://nyc.gov/html/datamine	The City of New York
OpendataNI	http://www.opendatani.info	Northern Ireland, UK
Open Data Euskadi	http://opendata.euskadi.net	Basque Country, Spain
Pic and Mix	http://picandmix.org.uk	Kent County, UK
Toronto.ca/open	http://toronto.ca/open	The City of Toronto, Canada
Vancouver Open Data Catalogue	http://data.vancouver.ca	The City of Vancouver, Canada
Warwickshire Open Data	http://opendata.warwickshire.gov.uk	County of Warwickshire, UK

2.4 Linked Open Data

The term Linked Data refers to *"data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets"* (Bizer et al., 2009a, p. 2). Linked data is based on Semantic Web philosophy and technologies but in contrast to the full-fledged Semantic Web vision, it is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inferencing (Hausenblas, 2009). It promises the creation of the "Web of data" as data from decentralized and heterogeneous sources can be interlinked through typed links. Web of data aims at replacing isolated data islands with a giant distributed dataset built on top of the Web architecture (Heath, 2008).

Linked Data following a RESTful approach requires the identification of resources with URI references that can be dereferenced over the HTTP protocol into RDF data that

describes the identified resource. In addition, Linked Data include the creation of typed links between URI references, so that one can discover more data. More specifically, the four Linked Data principles as described by Berners-Lee (2010) are the following:

- All item should be identified using URIs;
- All URIs should be dereferenceable, that is, using HTTP URIs allows looking up the item identified through the URI;
- When looking up a URI it leads to more data, which is usually referred to as the follow your nose principle;
- Links to other URIs should be included in order to enable the discovery of more data.

Linked data distinguishes between information and non-information resources. The former refers to all the resources we find on the traditional document Web such as documents, images etc, while the latter refers to real world thing such as people, schools, laws, public agencies etc. The adoption of identifiers ensures to uniquely identify information resources in the Web but not the real world things the information resources refer to. Hence a central issue in the Web of data is the finding of identifiers that refer to the same real world thing. These identifiers became known as URI aliases.

The use of Linked Data technologies for publishing data on the Web provides the following advantages:

- Enables data to be integrated with the Web . This describes the ability to link together different pieces of information published on the Web and the ability to directly reference a specific piece of information.
- Reduces the challenge of integrating heterogeneous data and building large-scale, ad hoc mashups (Hausenblas, 2009; Heath, 2008).

The specification of the Linked Data principles Berners-Lee (2010) resulted in the emergence of the Web of Linked Data, which currently comprises more than 1000 datasets in various domains Schmachtenberg et al. (2014). The Linking Open Data (LOD) cloud diagram depicts the Web of Linked Data and aims at showing datasets that have been

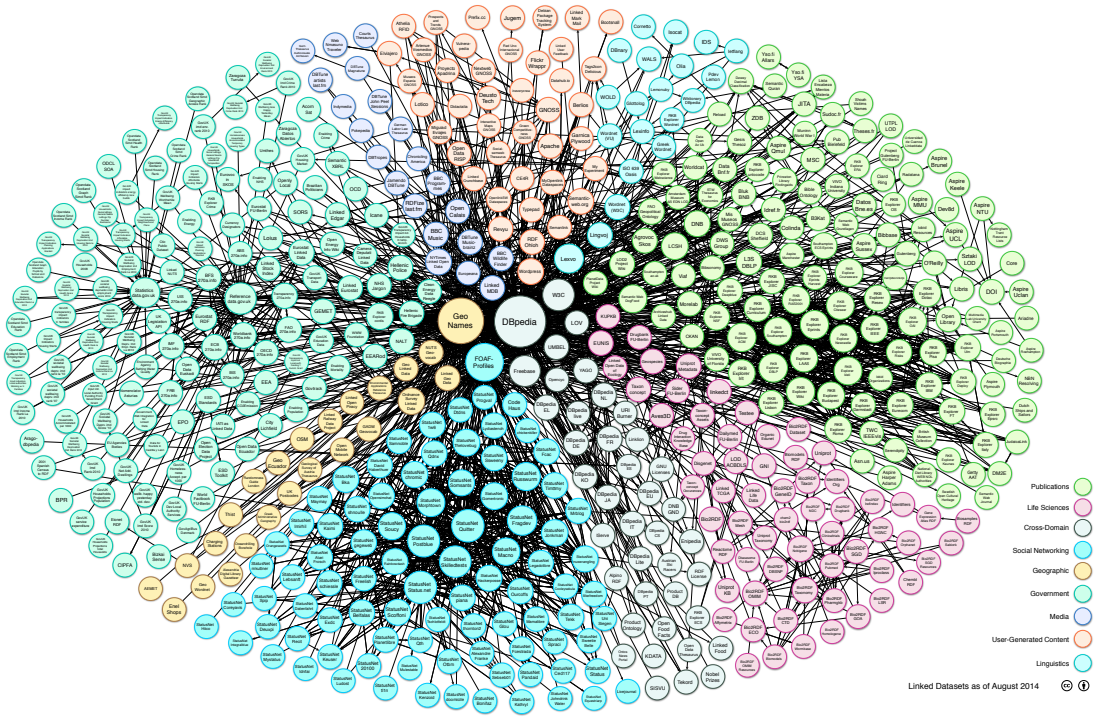


FIGURE 2.1: Linking Open Data cloud diagram, by R. Cyganiak and A. Jentzsch (2014 update) http://lod-cloud.net/

published in Linked Data format and are linked to other datasets. Figure 2.1 presents the LOD cloud diagram, where the different datasets are depicted as bubbles and the connections between datasets as arrows. The direction of the arrows indicate the dataset that contains the links, e.g., an arrow from A to B means that dataset A contains RDF triples that use identifiers from B. Bidirectional arrows usually indicate that the links are mirrored in both datasets.

The DBpedia knowledge base (Bizer et al., 2009b) lies at the heart of the LOD cloud. The DBpedia project is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web. As DBpedia covers a wide range of domains and has a high degree of conceptual overlap with other datasets of the LOD cloud, an increasing number of data publishers have started to set RDF links from their data sources to DBpedia, making DBpedia one of the central interlinking hubs of the LOD cloud.

2.5 Multidimensional Open and Linked Data

A major part of open data concerns statistics such as demographics and economic indicators. For example, the vast majority of the datasets published on the open data portal of the European Commission[7]are of statistical nature. Major providers of statistics at the international level include Eurostat[8], World Bank[9], OECD[10], and CIA's World Factbook[11].

Statistical data is often organised in a multidimensional manner where a measured fact is described based on a number of dimensions, e.g. unemployment rate could be described based on geographic area, time and gender. In this case, statistical data is compared to a cube, where each cell contains a measure or a set of measures, and thus we onwards refer to statistical multidimensional data as data cubes or just cubes.

A data cube is specified by a set of dimensions and a set of measures. The dimensions create a structure that comprises a number of cells, while each cell includes a numeric value for each measure of the cube. Let us consider as an example a cube from Eurostat with three dimensions, namely time in years, geography in countries, and age group, that measures the employment rate. An example of a cell in this cube would define that the percentage of unemployed people between 25 and 49 years old in France in 1999 is 10.2 % (Figure 2.2).

The multidimensional data model, which is often compared to a data cube, was introduced to define the analytic requirements of *Online Analytical Processing (OLAP)* and *data warehouse (DW)* systems. The notion of OLAP that were introduced by Codd et al. (1993), refers to the technique of performing complex analysis over the information stored in a DW. A DW is a large data repository with integrated historical data organised specifically for analytical purposes. Inmon (2005) defined a DW as a collection of subject-oriented, integrated, non-volatile, and time-variant data to support management's decisions.

Although research in OLAP and DW is active for more than two decades, concepts and systems lack a uniform theoretical basis with regards to models that define data

[7] http://open-data.europa.eu
[8] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes
[9] http://data.worldbank.org
[10] http://www.oecd.org/statistics/
[11] https://www.cia.gov/library/publications/the-world-factbook/index.html

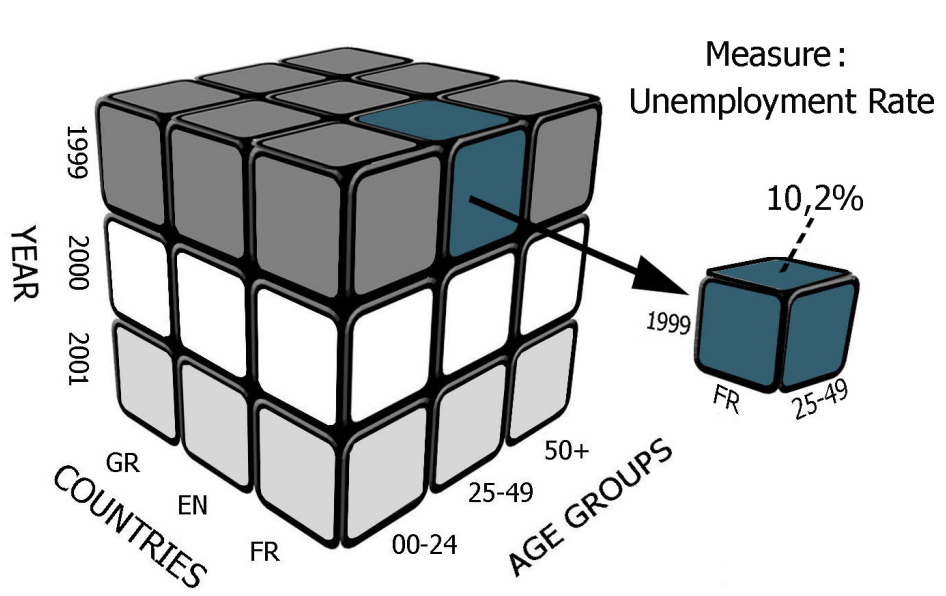


FIGURE 2.2: Aggregated multi-dimensional data modelled as a cube

cubes and operations that are performed on data cubes (Chaudhuri and Dayal, 1997; Datta and Thomas, 1999; Niemi et al., 2003; Pedersen et al., 2001; Ravat et al., 2008; Romero and Abelló, 2009; Tseng and Chen, 2005; Vassiliadis, 1998). In order to define the semantics that are related to multidimensional models we study relevant terminology proposed in the literature.

In general, as described in Romero and Abelló (2009), *dimensional concepts* structure the multidimensional space where the fact is placed. Dimensional concepts can be used as a perspective of analysis and have been classified as *dimensions*, *levels* and *descriptors*. A dimension is considered to contain a *hierarchy of levels* representing different granularities (or levels of detail) to study data, and a level to contain *descriptors*. On the other hand, a *fact* contains *measures* of analysis. One fact and several dimensions to analyze it give rise to a *multidimensional schema*. Finally, *base* is a minimal set of levels functionally determining a fact. Thus, two different instances of data cannot be placed in the same point of the multidimensional space.

More specifically, in Datta and Thomas (1999) an abstract structure of a data cube is defined as a 4–tuple, $< D, M, A, f >$ where D is a set of n dimensions, M is a set of k measures, A is a set of t attributes (or levels), and f is a one-to-many mapping from each dimension to a set of attributes $f : D \rightarrow A$. A cube instance of this abstract structure is defined by a 6–tuple $< D, M, A, f, V, g >$ where the elements D, M, A, f are

inherited from the "parent" cube while V represents a set of values that have been used to materialise the cube and g represents a mapping of the values to the specific cells of the cube.

Moreover, in Ravat et al. (2008) a constellation C_S is defined as a broader than a cube concept that comprises a set of facts, a set of dimensions and a function that associates each fact to its linked dimensions. A dimension is defined by a set of attributes, a set of hierarchies and a set of dimension instances. Each attribute represents one data granularity according to which measures could be analysed. A hierarchy of a dimension is defined by an ordered set of attributes and a function associating each parameter to one or several attributive properties. Finally, a fact is defined by a set of measures associated with and aggregate function, a set of fact instances, and a function that associates fact instances to their linked dimension instances.

In Tseng and Chen (2005) a dimension is defined as a tree structure that represents the hierarchical relationships among a set of members. The i-th level members set defines the i-th level of D, while an ordered set of levels defines the hierarchy of D. A cell $c = (t_c, M_c)$ is defined in n dimensions $(D_1, D_2, \ldots, D_n)$where t_c is a set that contains a member for each one of the n dimensions, while M_c is a tuple consisting of measures. A data cube $C = (M, D_1, D_2, \ldots, D_n)$ is a cube composed of all cells $c_i = (t_{c_i}, M_{c_i})$ with $t_{c_i} \in \times_{1 \leq j \leq n} D_j(0)$, where $M = \bigcup_i \{M_{c_i}\}$ is the measure dimension of C.

2.5.1 Linked Data Cubes

The RDF Data Cube (QB) vocabulary Cyganiak and Reynolds (2014) is a *W3C* standard for modelling data cubes as graphs and thus adhering to the RDF model and Linked Data principles. Centric class in the vocabulary is *qb:DataSet* that defines a cube. A cube has a *qb:DataStructureDefinition* that defines the structure of the cube and multiple *qb:Observation* that describe each cell of the cube. The structure is specified by the abstract *qb:ComponentProperty* class, which has three sub-classes, namely *qb:DimensionProperty*, *qb:MeasureProperty*, and *qb:AttributeProperty*. The first one defines the dimensions of the cube, the second the measured variables, while the third structural metadata such as the unit of measurement.

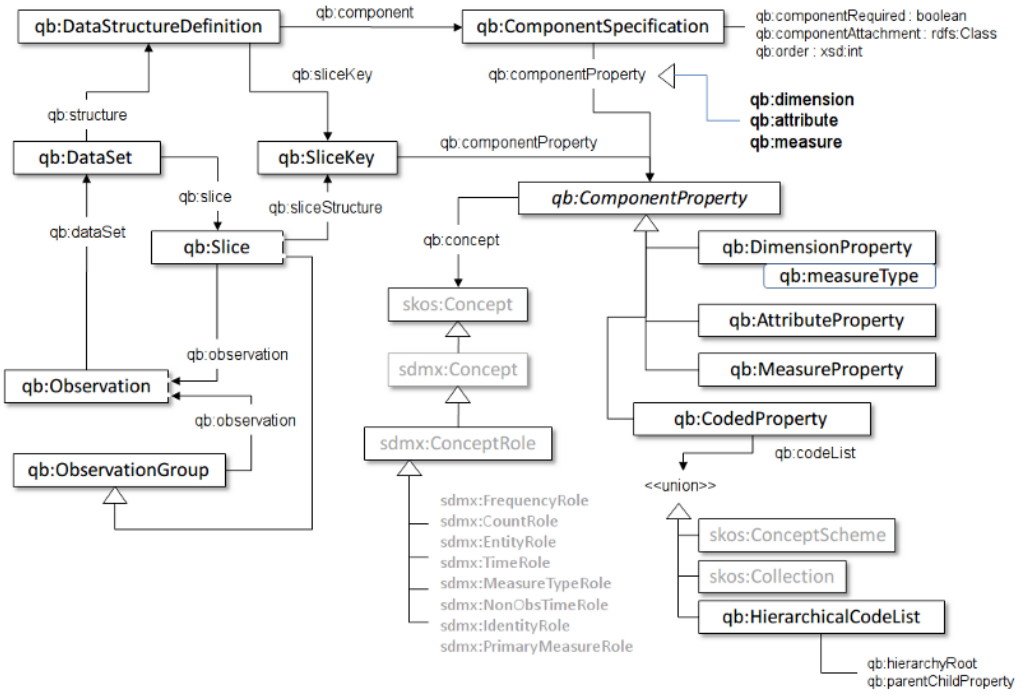


FIGURE 2.3: The RDF Data Cube Vocabulary (Cyganiak and Reynolds, 2014)

Usually the values of the components are populated using predefined code lists that might formulate hierarchies such as a geographic or administrative division. These code lists can be specified by using either the Simple Knowledge Organization System (SKOS) (Miles and Bechhofer, 2009) vocabulary or the QB vocabulary. SKOS is a W3C standard used for expressing the basic structure and content of concept schemes such as thesauri, taxonomies, and classification schemes. The set of values is modelled as a *skos:ConceptScheme* and a value as a *skos:Concept*. In addition, *skos:broader* and *skos:narrower* are used to assert a direct hierarchical link between two *skos:Concepts*. In case of reusing RDF data that are not modelled using SKOS, the QB vocabulary introduced the *qb:HierarchicalCodeList* class that defines a set of root concepts in the hierarchy (*qb:hierarchyRoot*) and a parent-to-child relationship (*qb:parentChildProperty*).

XKOS[12] RDF vocabulary has been proposed as an extension to SKOS that allows to model hierarchies structured in levels. A hierarchy level can be defined using the *xkos:ClassificationLevel* concept. According to XKOS the levels of a hierarchy are organised as an *rdf:List*, which implies order, starting with the most aggregated level. Individual *skos:Concept* objects are related to the *xkos:ClassificationLevel* to which they belong by the *skos:member* property.

[12] http://rdf-vocabulary.ddialliance.org/xkos

At the moment, a number of statistical datasets are freely available on the Web as linked data cubes. For example, the European Commission?s Digital Agenda provides its Scoreboard as linked data cubes. An unofficial linked data transformation of Eurostat's data[13], created in the course of a research project, includes more than 5,000 linked data cubes. Few statistical datasets from the European Central Bank, World Bank, UNESCO and other international organisations have been also transformed to linked data in a third party activity[14]. Census data of 2011 from Ireland and Greece and historical censuses from the Netherlands have been also published as linked data cubes (Meroño-Peñuela et al., 2012; Petrou et al., 2013). Finally, the Department for Communities and Local Government (DCLG) in the UK also provides local statistics as linked data[15].

The real value, however, of linked data cubes is revealed in the case of combining statistics from disparate sources and performing analytics on top of them in an easy way. Let us consider a cube from Eurostat that measures the population involved in life long learning and a cube from Digital Agenda that measures internet usage. Both cubes are structured based on the same three dimensions, i.e. time in years, countries, and sex. If we combine these two cubes from Eurostat and Digital Agenda, we can perform a regression analysis and derive some interesting results like the plot of Figure 2.4. In this case, the value is present when all needed steps including discovery, integration, and exploitation can be easily performed using relevant online tools.

2.6 Other Web Data

2.6.1 Social Media Data

In the last years, Social Media (SM) have grown in popularity with millions of users voluntary submitting huge amounts of data in various forms such as text messages, tags and multimedia content. SM data incorporates personal opinions, thoughts and behaviours making it a vital component of the Web and a fertile ground for a variety of business and research endeavours.

[13] http://eurostat.linked-statistics.org
[14] http://270a.info
[15] http://opendatacommunities.org

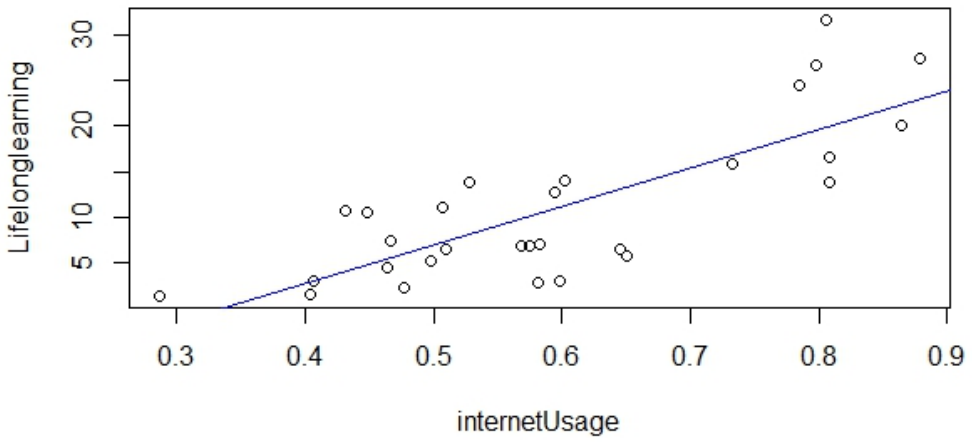


FIGURE 2.4: Combining and analysing data cubes from Eurostat and Digital Agenda (Tambouris et al., 2015)

Twitter is one of the most popular Social Media with 288 million monthly active users publishing more than 500 million tweets per day (as of January 2015). This is mainly due to the open nature of Twitter API.

While most tweets are conversation and chatter, many users share information and spread news through Twitter as well (Java et al., 2007; Naaman et al., 2010). The number of tweets that are posted online is closely related to real world phenomena, as important events seem to trigger an increased number of tweets (Hughes and Palen, 2009). It is interestingly estimated that the majority of the trending topics and 85% of all tweets are related to news (Kwak et al., 2010). In Twitter realm traditional media organisations seem to play a significant role since they are by far the most active users. However, only about 15% of tweets received by ordinary users are received directly from the media (Wu et al., 2011).

Two important features of Twitter are hashtags and the symbol. A hashtag is a character string preceded by the # symbol and denotes some aspects of a tweet such as its topic or its intended audience. The symbol is mainly used to address a tweet to a particular user and thus enables the identification of conversations in Twitter. The symbol is also used to mark a retweet using either “RT” followed by “ user id” or “via

@user id". Retweets enable the forwarding of a tweet by posting it again, sometimes with additional comments (Kwak et al., 2010).

A characteristic of Twitter that differentiates it from other popular Social Media is that the most influential users are not determined by the number of followers or the number of their posts. It has been experimentally proved that the most important mechanism that determines the power of a tweet to spread information broadly is retweets (Kwak et al., 2010). So, no matter how many followers a user has, the tweet is likely to reach a certain number of audience, once the user's tweet starts spreading via retweets.

As Twitter gains popularity and becomes a valuable media to spread information, illegitimate use such as spam emerges as well (Grier et al., 2010; Wang, 2010). In the political domain for example, spam in Twitter is defined as campaigns disguised as spontaneous, popular grassroots behaviour that are in reality carried out by a single person or organization (Ratkiewicz et al., 2011). Castillo et al. (2011) suggested that credible news are propagated in Twitter through authors that have previously written a large number of messages, originate at a single or a few users in the network, and have many re-posts.

Twitter data has been recently considered as a source of data that if published as Linked Data and connected to other data sets would provide value to the users. It was suggested that Linked Data could alleviate the information overload problem of Twitter by replacing keyword and hashtag based search by SPARQL queries, which would specify RDF graph patterns as constraints and thus users could select a subset of data that matches their needs (Mendes et al., 2010). In the same context, an architecture and a tweet format that would enable the classification of tweets through semantically enriched hashtags were proposed (Shinavier, 2010). Moreover, Rowe & Stankovic (Rowe and Stankovic, 2012) proposed an approach to automatically align tweets with events, while Abel et al. (Abel et al., 2011) proposed a semantically enriched faceted search method for Twitter, which is based on Linked Data paradigm.

In many cases in Twitter, vital pieces of information are provided not within the text of the post, but behind hyperlinks that users post to refer to a relevant resource. So, these hyperlinks have also been used as pointers to additional Linked Data that is related to the specific tweet. In particular, Kinsella et al. (2011) proved that this sort of information

improves the classification of Twitter posts by implementing a multinomial Naive Bayes classifier.

Named Entity Recognition (NER) is the process of identifying named entities in text and classifying them into predefined categories such as person, organization and location (Nadeau and Sekine, 2007). NER is very important in Linked Data representation of tweets as it enables the identification of named entities included in the actual text and thus it enriches the actual representation. NER methods can be classified into two main categories: Rule Based Methods and Machine Learning Methods. The former includes a set of rules that implements a specific grammar to identify and classify named entities while the later is divided into three categories i.e. supervised learning (SL), semi-supervised learning (SSL) and unsupervised learning (UL). In SL the classifier for NER is trained using a training data set that is annotated with named entities while in UL there is no annotated data.

Due to the informal nature and short length of tweets, the performance of conventional text analysis tools drops sharply in Twitter (Finin et al., 2010; Liu et al., 2011). In order to tackle these challenges different approaches have been proposed in the literature. Some studies aim at syntactical and lexical normalising ill-formed words and expressions usually found in tweets (Han and Baldwin, 2011; Kaufmann and Kalita, 2010). Some other works propose new statistical methods that could overcome the limitation of existing tools based on Twitter's requirements. In this context, (Liu et al., 2011) proposed a combination of K-Nearest Neighbors (KNN) based classifier and a linear Conditional Random Fields (CRF) based labeler under a semi-supervised learning framework.

2.6.2 Scientific Data

Sharing scientific data through publication has long underpinned the cycle of discovery. However, many scientific manuscripts consist of a "full–text" manuscript and additional "supplemental data" or "supporting information" (Murray-Rust, 2008). Example of such scientific data include data and resources from hypothesis-driven research or from predictive analyses. This data is important (Murray-Rust, 2008; Schofield et al., 2009):

- for the validation of published results, scientific data could supplement a scientific manuscript because it is necessary for a reader who wants to be sure that the

experiment has been carried out correctly and that the right conclusions have been drawn, and

- for reuse in other purposes e.g. classic example of reuse is Mendeleev's use of published data to propose the Periodic Table of the Chemical Elements. In this case the published data (melting points, colours, densities, etc.) were often not collected for a specific purpose other than the worthy belief that data was, per se, valuable

In general shared, annotated research data and methods offer new discovery opportunities and prevent unnecessary repetition of work and thus the need to open up scientific data emerges (Molloy, 2011; Sansone et al., 2012; Schofield et al., 2009).

2.6.2.1 Clinical Trial Data

Clinical research aims at finding new and better ways to understand, diagnose, prevent, or treat a specific pathological process, e.g., diseases or adverse events. It comprises three main categories: (i) the patient-oriented research that involves human subjects; (ii) the epidemiological and behavioral studies that examine the distribution of disease and the factors that affect health; and (iii) the outcomes and health services research that seeks to identify the most effective and efficient interventions, treatments, and services.

Clinical research often requires the integration of medical data coming from multiple datasets that are usually stored across multiple sources such as hospitals, clinical sites, research institutes and pharmaceutical companies (Burgun and Bodenreider, 2008; Maro et al., 2009; Prokosch and Ganslandt, 2009; Weiner and Embi, 2009). Medical data may contain sensitive patient data such as demographics, diagnoses, and medication, as well as radiology images, laboratory test results, doctors entries and comments (Jensen et al., 2012; Lau et al., 2011).

In patient-oriented research, the integration of multiple medical datasets enables the identification of a sufficient number of subjects (Anderson et al., 2012). For example, clinical trial phase III, which assesses the safety and the efficacy of a studied treatment or drug, requires large groups of people matching specific eligibility criteria that cannot be found through a single clinical site. In epidemiological studies, analysis of integrated

datasets improves the statistical power of results. For instance, studies of clinical effectiveness or disease biology in rare diseases are only possible through multi-center analyses (Sherborne et al., 2011). The integration of multiple datasets also enables better understanding of relationships between pathological processes and risk factors, or between genotype and phenotype (McMurry et al., 2013; Tong and Zhao, 2008). For example, recent genome-wide association studies identified 13 novel loci associated with systolic and diastolic blood pressure as well as hypertension (Levy et al., 2009; Newton-Cheh et al., 2009).

At the same time however, clinical researchers face technical and interoperability (Kush et al., 2008), as well as ethical and legal (Taylor, 2008), challenges in discovering and accessing scattered and heterogeneous medical data. Although the former challenges have been addressed by several standards (Begoyan, 2007; Goossen et al., 2010), the latter still remain.

In order to overcome these ethical and legal challenges, the approach of aggregating data has been proposed and widely employed. According to this approach, only the counts of subjects having specific characteristics are reported instead of raw record-level data, guaranteeing in this way non-identification and anonymisation. Despite that, there is still need for controlling access even to aggregated data, e.g., due to data providers policies.

The importance of publishing medical data as Linked Data becomes apparent in the case that we reuse widely used ontologies or linked datasets. For example, the International Classification of Diseases-11 ontology (Tudorache et al., 2013) classifies diseases and other health problems, including signs, symptoms, abnormal findings, etc. In addition, the Experimental Factor Ontology (EFO) (Malone et al., 2010) combines parts of several biological ontologies, such as anatomy, disease and chemical compounds, to annotate experimental variables. Furthermore, an increasing amount of life science datasets are becoming available as linked data. UniProt is a database of protein sequence (Consortium, 2013), Reactome describes biological pathways (Croft et al., 2011), ChEMBL contains bioactive drug-like small molecules (Willighagen et al., 2013), ChemSpider includes chemical compounds (Pence and Williams, 2010), and WikiPathways describes pathways (Kelder et al., 2012). The reuse of such ontologies and linked datasets enables

the disambiguation of concepts referring to the same entity, as well as the enrichment of medical data with data coming from disparate sources.

The RDF Data Cube vocabulary is also of vital importance in Linked Medical Data because it can be used to publish the aggregated data that clinical researchers produce.

Chapter 3

Provide Open Data

3.1 Introduction

This chapter presents the OGD framework that provides a holistic understanding (both technical and organisational) of OGD. The framework comprises two conceptual models:

- A classification scheme for OGD (section 3.2): The classification scheme is based on relevant literature and enables identifying, analysing and classifying OGD initiatives. The classification scheme comprises two main dimensions. The first dimension refers to the technological approach followed for making data available on the Web, while the second refers to the organisational approach followed for providing OGD. We believe that having a classification scheme allows a deeper understanding of initiatives and therefore the domain as a whole.
- A stage model for OGD (section 3.3): The stage model aims at (a) providing a roadmap for open government data re-use and (b) enabling evaluation of relevant initiatives' sophistication. The proposed model has two main dimensions, namely organisational & technological complexity and added value for data consumers.

3.2 Classification Scheme

A significant part of eGovernment initiatives involves the development of governmental portals acting as single point of access to governments. One-stop government portals face

amongst others legal, organisational, technological and cultural challenges (Tambouris, 2001). In this conceptual analysis, we concentrate on organisational aspects arising from the establishment of a portal acting as a single point of access as opposed to making data available from each public agency's website.

For our analysis, we employ as a starting point the *data manufacturing systems* paradigm where the production and storage of data has been conceptualised (Strong et al., 1997). In data manufacturing systems, three roles are identified and each role is associated with a process:

- data producers are associated with data production process;
- data custodians with data storage, maintenance and security; and
- data consumers with data utilisation processes, which may involve additional data aggregation and integration.

In the case of OGD initiatives the three roles become:

- R1 customer (which is actually the data consumer)
- R2 one-stop government data portal (portal)
- R3 public agency (which is actually the data producer).

Furthermore, the main data-related processes are:

- P1 data production process
- P2 data utilisation processes
- P3 data publishing and maintenance process
- P4 data aggregation and integration process
- P5 data searching process
- P6 data collection process.

Based on the main activities required for publishing linked data, we derive three more data-related processes, which are present in OGD initiatives adopting a linked data approach:

- P7. URI definition and management process;
- P8. Data Vocabulary creation and management; and
- P9. Data Links creation and management.

Hence, OGD activities involve three main roles and nine processes relevant to data and metadata management. It is important to note that unlike the data manufacturing systems paradigm where each process is assigned to one actor, in OGD initiatives more options exist and the decisions might have wider consequences. The main processes and relevant options include:

- Who own (i.e., maintains) the data? This could be either the public agency or the portal. This is particularly important as it relates directly to data quality, e.g., data may become obsolete is not properly and timely maintained.
- Who publishes the data (and possibly related metadata)? This could be either the public agency or the portal or both. Again, this decision might have consequences in data quality. In case both public agencies and the portal publish data, it is possible that different values appear in different places due to lack of proper synchronisation.

It should be also noted that the above-mentioned decisions can be driven by different factors. An important factor is efficiency and effectiveness. Other factors include the principles of subsidiarity (which suggests that matters ought to be handled by the smallest, lowest or least centralised competent authority), legitimacy, transparency, accountability and trust. These are particularly important considerations in the case of public sector and therefore should be also considered in the case of OGD initiatives.

Finally, we should note that other challenges also exist when designing OGD initiatives. For example, a main challenge associated with establishing a main portal is which agency will host this portal. Other considerations include the shift of power which might be present when there is a change in the authority that own public data. These are both

strategic decisions but are outside the scope of this thesis hence will not be further considered.

The proposed classification scheme includes two dimensions. The first dimension cares for the technological aspect of OGD initiatives, which is an important driver. In Chapter 2, we have already presented purely technology-driven schemes that can be used for classifying OGD initiatives (although it should be admitted that they were not initially proposed as classification schemes). The second dimension cares for the fact that OGD activities are actually online one-stop government data portals thus non-technical, domain-specific peculiarities should be also considered. In summary, the proposed classification scheme comprises the following dimensions:

- the technological approach followed for making data available on the web
- the organisational approach followed for data provision.

Each dimension is elaborated before presenting the proposed classification scheme.

The first dimension refers to the technological approach followed. In general there are many different technological approaches for making data available on the web such as:

- as downloadable files in proprietary formats
- through custom APIs
- as downloadable files in machine readable formats
- through RESTful APIs
- through search interfaces.

These are characterised by how easy it is to

- use the relevant technology,
- access the data over the web,
- extract and reuse data, and
- link together different pieces of information.

The proposed scheme includes two broad technological approaches. These are further divided according to Tim Berners-Lee five-star technological maturity model.

- The first approach suggests making data available on the web as downloadable files in well-known formats such as PDF, Excel, CSV, KML, XML, JSON etc. This broad category is further divided based on the format in:
  - making data available on the web as downloadable files in proprietary and desktop-centric formats, e.g., PDF.
  - making data available on the web as downloadable files in machine-readable formats, e.g., Excel.
  - making data available on the web as downloadable files in machine-readable formats using open standards, e.g., CSV.
- The second approach suggests making data available on the web as linked data through RESTful APIs and/or SPARQL search interfaces. This broad category is further divided in:
  - making data available based on Linked Data principles (i.e., HTTP, URIs and RDF).
  - linking data from different datasets.

The second dimension of the proposed classification scheme is related to the organisational approach followed for providing governmental data. In the motivation of this thesis (Chapter 1), we reviewed several e-government models that can be used as classification schemes while in this chapter, we highlighted the challenges of OGD initiatives. Based on this analysis, we conclude that an adequate classification of ODG initiatives should include the organisational approach they are following for providing data. Broadly speaking, two approaches are possible:

- Data belonging to various public agencies is published by the one-stop government data portal. We use the term direct data provision to dictate this method of data provision.
- Data belonging to various public agencies is published in a decentralised manner by these agencies (usually in their website) while the portal provides some kind

of linking mechanism and/or metadata for the identification of the actual dataset. We use the term indirect data provision to dictate this method of data provision.

Clearly, other approaches can be also considered for covering the organisational aspects of OGD initiatives. We believe however that selecting the data provision approach enables us to address most of the public sector considerations in a simple and straightforward manner.

By combining the technological and organisational dimensions we derive the classification scheme shown in Figure 3.1.

<table>
<tr><th colspan="2">Technological Approach \ Organizational Approach</th><th>Direct Data Provision</th><th>Indirect Data Provision</th></tr>
<tr><td rowspan="3">Downloadable Files</td><td>Proprietary and desktop-centric formats</td><td rowspan="3">Repository of Downloadable Files</td><td rowspan="3">Registry of Downloadable Files</td></tr>
<tr><td>Machine-readable formats</td></tr>
<tr><td>Machine-readable formats using open standards</td></tr>
<tr><td rowspan="2">Linked Data</td><td>Linked data principles</td><td rowspan="2">Direct Provision of Linked Data</td><td rowspan="2">Indirect Provision of Linked Data</td></tr>
<tr><td>Linking available data</td></tr>
</table>

FIGURE 3.1: The proposed OGD classification scheme (Kalampokis et al., 2011b)

We now outline the main characteristics of each of the four main dimensions:

- Downloadable files: The main advantage is that data is provided in simple to use formats that are widely accepted by both developers and customers, e.g., citizens and businesses.
- Linked data: The use of linked data technologies infuses the technical advantages of linked data, i.e., ability to link to a specific piece of data and reusing part of the data. On the other hand, the effort and time needed for publishing the data (i.e., finding vocabularies, assign URIs etc.) are large while at the same time the relevant technologies are still immature and not widely adopted. Furthermore,

technological challenges still exist such as those related to standardised querying of distributed data sources using SPARQL.

- Direct data provision: Having all data at one place (portal), suggests that aggregated, processed and value-added data can be provided by the governmental portal. On the other hand, maintainability is limited, e.g., in cases where data change over time, an efficient and effective data synchronisation process must be in place to prevent the portal from providing obsolete data.

- Indirect data provision: The fact that the actual data is published by the data producer itself means that the provided data is the only one of its kind (unique) and also up to date. These characteristics contribute to the increase of data believability and data accuracy. On the other hand, aggregated, processed and value data cannot be provided by the portal; if this is needed, it has to be performed by the customer.

As already suggested, we have identified three main roles (customer, portal, public agency) and nine different data-related processes. Figure 3.2 shows the main data-related processes in each class of the proposed scheme, namely repository of downloadable files, registry of downloadable files, direct provision of linked data, and indirect provision of linked data. Figure 3.2 clearly illustrates that the same data process can be performed by different actors.

In Figure 3.3, the identified OGD initiatives are classified according to the proposed scheme. The majority of the initiatives fall in the first class (i.e., direct data provision based on downloadable files) while the third class (i.e., centralised provision of linked data) includes only three initiatives. In addition, there is only one initiative (i.e., data.gov.uk) falling in the fourth class characterised by the indirect provision of linked data. We should also note that a number of initiatives fall into more than one categories. This is due to the fact that these initiatives use more than one technological approach. For example they provide data in both proprietary and open formats. However, we should underline that the organisational approach followed is always the same with data.gov.uk being the only exception.

Data.gov.uk includes the indirect provision of data in machine-readable formats using both proprietary and open formats as well as the provision of linked data in both a direct

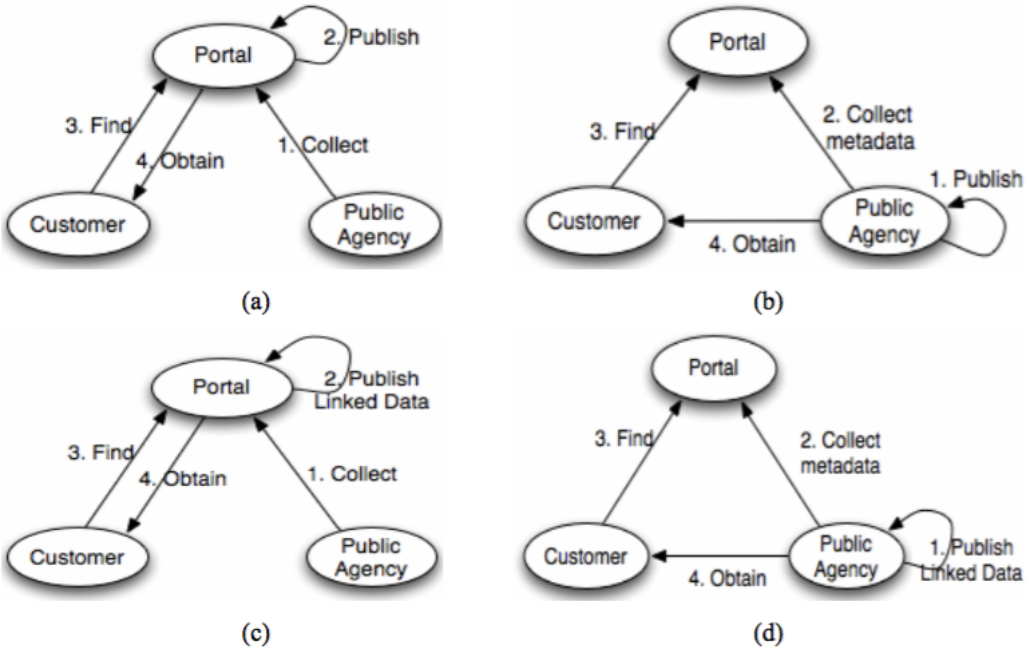


FIGURE 3.2: Main data processes in the classes of the proposed scheme (Kalampokis et al., 2011b)

and indirect manner. In our understanding, the provision of linked data by different public agencies such as the HM Treasury, the Department of Education and the Office for National Statistics using sub-domains of data.gov.uk (e.g., education.data.gov.uk) denotes direct data provision. But we consider the provision of geo-spatial data by ordnance survey as indirect provision of linked data because it is performed in the agency's official URL.

Data.gov.uk is the only initiative included in the Linking available data sub-class as links between different datasets have been created. In Figure 3.4, an example of this linking is depicted. More specifically, data from the Department of Education describing schools is linked to data from the Office for National Statistics. The joint point of these datasets is the Local Learning Skills Council (LLSC) that is responsible for the specific school.

The analysis suggests only one OGD initiative (namely data.gov.uk) adopts linked data and a partially indirect data provision approach. We have already shown that an indirect data provision approach features some interesting public sector characteristics (trust, accountability etc) and that linked data as a technology can fully support it but poses additional considerations. However, we could not find an OGD initiative that

Technological Approach \ Organizational Approach		Direct Data Provision	Indirect Data Provision
Downloadable Files	Proprietary and desktop-centric formats		Data.govt.nz Data.nsw Mass.gov/data Open Data Euskadi
	Machine-readable formats	NUC data mine Pic and Mix Toronto.ca/opne Vancouver's open data catalogu	Data.australia.gov.au Data.ca.gov Data.gov.uk Data.govt.nz Data.nsw Data.vic.gov.au DataSF OpendataNI
	Machine-readable formats using open standards	City of Edmonton open data catalogue Data.gov Data.seattle.gov Dati.piemonte.it Dc.gov - data catalogue Lichfield open data London datastore Maine.gov DataShare Toronto.ca/open Vancouver's open data catalogue Warwickshire open data	Data.ca.gov Data.gov.uk Data.vic.gov.au
Linked Data	Linked data principles	Catalogo de Datos de Asturias Data.gov Data.gov.uk	
	Linking available data		Data.gov.uk

FIGURE 3.3: OGD initiatives grouped according to the classification scheme (Kalampokis et al., 2011b)

fully adopts this operation model. Chapter 4 elaborates on this and proposes architectures and theory for discovering and integrating compatible OGD that are provided by different public agencies at multiple administrative levels.

3.3 Stage Model

Based on our previous analysis we now describe a stage model aiming at proposing a number of steps that OGD should go through in order to generate new integrated datasets that increase the added value that might potentially produce by their further

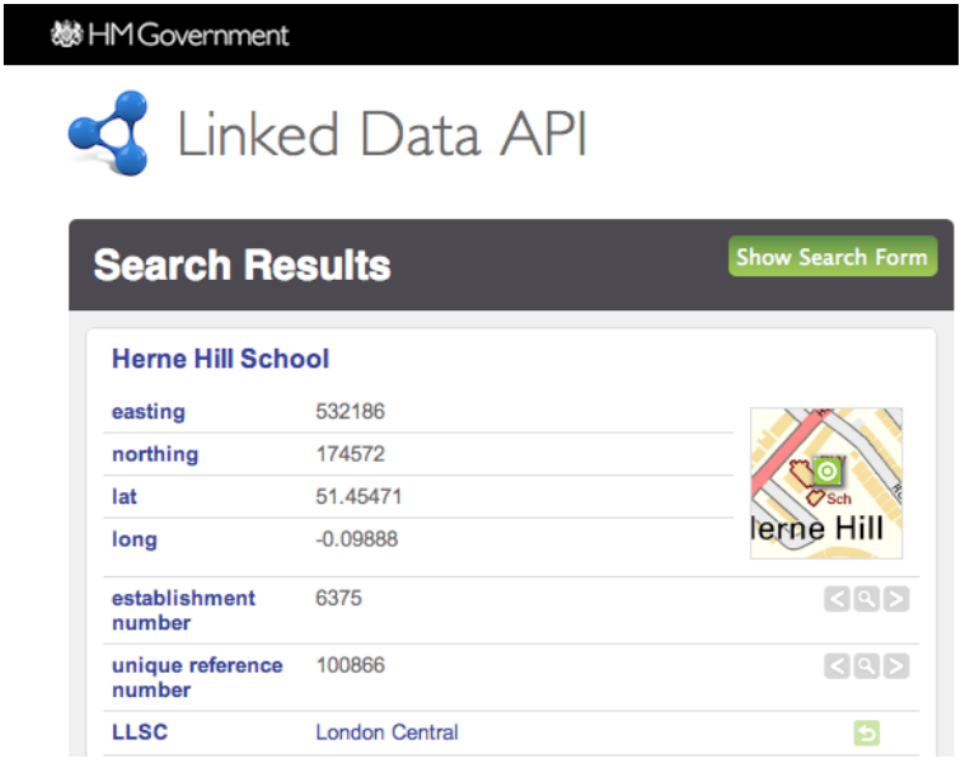


FIGURE 3.4: Screen-shot of data.gov.uk site showing the linking of data from the Department of Education (schools) to data from the Statistics Office (admin areas) (Kalampokis et al., 2011b)

analysis. The proposed stage model comprises four stages as depicted in Figure 3.5. The first two stages deal with OGD integration inside public administration while the last two stages with the integration of OGD with the other non-government data on the Web.

The model capitalises on stage models proposed to measure the development of eGovernment (see section 1.4). Unlike these models, however, where focus is on service provision the proposed model's focus is on *data integration*.

In Figure 3.5, the vertical axis presents the technological and organisation complexity that is involved in the provision of the data while the horizontal axis presents the capability of developing added value services based on the provided data. In this section, we describe the four stages of the proposed model.

The aim of the model is two-fold: first to provide a roadmap for open government data re-use and second to enable evaluation of relevant initiatives' sophistication.

3.3.1 Stage 1: Aggregation of Government Data

This stage includes opening up data, publishing data online for others to re-use and, possibly, aggregating data provided by different sources. The main concern of public agencies in this stage is to easily and quickly make their data available online. Different agencies can publish their data employing different technological solutions and following different implementation details. This stage may also include data aggregation in a single website like the recently launched OGD portals. We use the term aggregation here to indicate that data is simply gathered and provided together from a single point of access.

In this stage, public agencies have to overcome a number of organisational, cultural and legislative barriers. In the case of European Union, the ultimate goal of the Directive on public sector information re-use is for all member States to overcome these barriers and hence provide their data online for anyone to re-use.

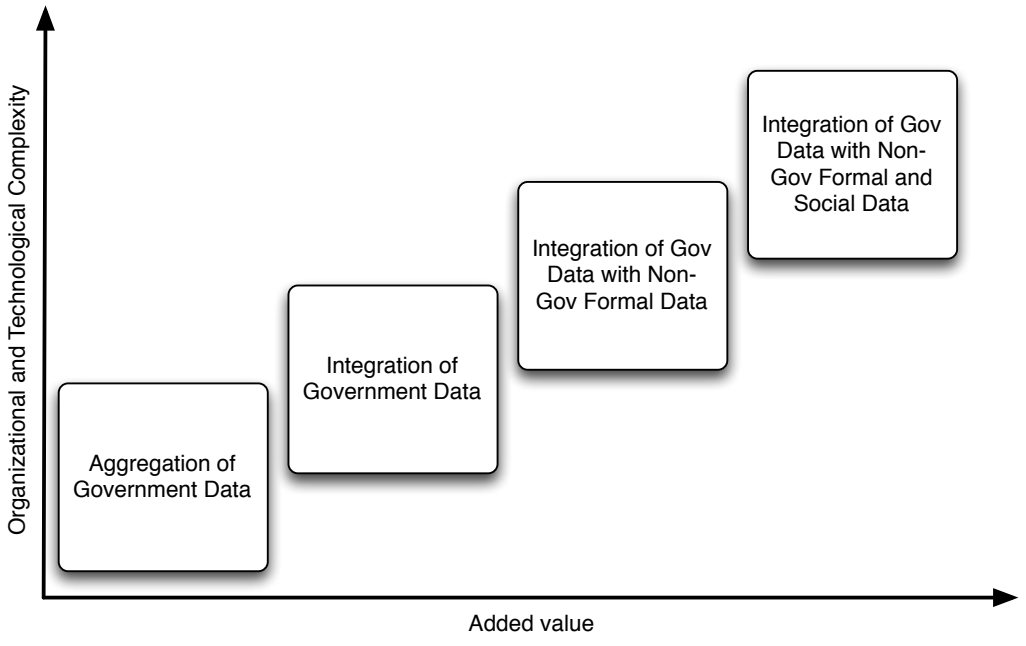


FIGURE 3.5: The Open Government Data Stage Model (Kalampokis et al., 2011c)

From an organizational perspective, open data at this stage can be provided in one of the following ways:

- The public agency publishes the data sets on its website or on the website of a higher-level agency.

- The public agency forwards the data set to an OGD portal that publishes the data.
- The public agency publishes the data sets and the OGD portal provides links to the actual data sets along with metadata.

From a technological perspective, the following approaches are possible according to the analysis presented in 3.2:

- Publish downloadable files in well-known formats such as CSV, XML, KML etc.
- Publish data using the linked data paradigm but without caring about linking to other data sets.

The main benefit of this first stage is that the public gains access to a wealth of valuable data. This data can be used for the development of new added value services. However, at this stage, governments do not consider a number of limitations that could impede data use and re-use. Actually, data is available as provided by agencies and thus it is not possible to automatically search across data provided by different agencies or combine them in order to create value-added services and products. According to the analysis presented in the motivation of the Thesis, these limitations are related to data duplication and data freshness, data formats that facilitate re-use, complete metadata, linking to other data sets etc.

As a result, at this stage data consumers need to be involved in a time and effort consuming process in order to overcome these limitations and use the provided data. This process could include the identification of all sources that provide data related to a specific real-world problem, assess the accuracy of the data, fuse the identified data sets, transform the data to the appropriate format, identify other data sets that could add value to the solution and integrate them with the initial ones.

Based on the analysis presented in section 3.2, in which a significant number of OGD initiatives were analysed, we deem that the majority of the existing OGD initiatives fall into this stage. An indicative example of government data provision in this stage can be given by Data.gov.uk where data on criminal statistics is provided both in the Ministry of Justice website and the central access point where in the latter an outdated version of the data exists.

3.3.2 Stage 2: Integration of Government Data

This stage includes government data integration across public administration. The analysis in the motivation of the Thesis presented a number of government data provision challenges that emerge from the decentralised structure of public administration. These challenges emerge when different agencies in different administrative levels and functional areas provide data about the same real-world problem since this data can be incomplete, controversial or obsolete.

The most important benefit of this stage is the provision of a unified view of government data that comes from different sources. In addition, it is expected that integrated government data will be complete and concise: complete suggests no specific object is forgotten; concise suggests no object is represented twice and data is without contradiction.

Government data integration is a very challenging task that includes significant technological and organisational issues. As regards the technological issues, governments should provide their data in specific formats that enable and facilitate integration on the Web. At the moment, Linked Data seems to be the most promising approach towards this direction. Thereafter, governments should decide on the architectural approach to follow (e.g. central repositories or federated queries). Other technological issues involved in this process are data schemas standardisation, identifiers standardisation, etc. With regards to the organisational issues, governments should establish business processes that prevent data re-publishing from different agencies, ensure in-time publishing and enhance data accuracy. Decentralised data provision could be a solution towards this direction i.e. every public agency to disseminate only the data that has the mandate to manage.

This type of integration will enable data consumers to execute more complex queries on top of the integrated data. A simple question that could be easily answered at this stage would be "Which governmental points of interest are located on a specific area?"

Although the final goal of this stage is to provide integrated government data across every public agency, it is more possible partial integration to take place in the beginning. These initial efforts can be developed around real-world objects or specific real-world problem related queries. We can deem that this is the case in Data.gov.uk where partial

integrated data is provided around specific real-world object such as schools, bus stops, members of parliament, geo-locations etc. In these initiatives linked data technologies are employed and also links have been established between data sets provided by different public agencies such as Ordnance Survey, the Ministry of Education and the London Gazette.

3.3.3 Stage 3: Integration of Gov Data with Non-Gov Formal Data

Government data can be characterised as formal as it is published by a highly trustworthy source. Data consumers assume that data published by governments is always accurate and reliable. However, many non-governmental sources also provide formal data on the Web in structured formats that allow for re-use. In this category we could encompass DBpedia , which is the linked data version of Wikipedia, and Data.nytimes, which is the New York Times linked open data set. Although the former is a social platform, users participation ensures that the provided information is objective, accurate and unbiased. This sort of sources provides data about real-world things such as organisations, people and locations as well as subject descriptors such as greenhouse gas emission.

The integration of government data with this non-government formal data defines the next stage of the proposed model. This type of integration will enable the provision of richer information to data consumers and will allow for more complex queries answering. A simple use case that will be enabled by this stage could include the identification of news posts that refer to public agencies or politicians connected to high expenditures in the governmental budget reports.

The implementation of this stage increases both organisational and technological complexity that should be overcome by governments and third parties. As regards the former, possible conceptual integration points between government and non-government formal data should be identified. These integration points will define use cases that could add value to data consumers. Thereafter, relevant government data sets and sources of non-government formal data should be identified and the required technological and organisational connections that will enable data integration should be established. Taking into account that Linked Data is the most advanced technological approach in government data provision, the technological requirements of this stage would be the establishment and maintenance of links between government and non-government data sets. In

addition, richer metadata should be included in order to describe these links and these data sources.

3.3.4 Stage 4: Integration of Gov Data with Non-Gov Formal and Social Data

The final stage of the proposed model covers the integration of government data with not only non-government formal data but also social data on the Web. We define social data as data that is created and voluntarily shared by citizens through social media platforms such as Twitter and Facebook. This sort of data is differentiated from government data and non-government formal data because it mainly communicates personal opinions, beliefs and preferences.

This type of integration will allow for new innovative services in which government data will provide a context of interpretation for social data. In particular, it will enable governments to consider citizens opinion expressed through social media in governmental decision-making processes; it will further allow citizens to deliberate in social media about public administration related real-world things such as laws and public agencies in a more explicit manner.

For example, at this stage governments and citizens will be able to answer questions such as "What is the opinion of citizens affected by a specific law about this law?" In addition, governments will be able to understand public sentiment on specific decisions by analysing integrated government and social data and thus take corrective actions that would alleviate the foreseen reactions.

Social data is streamed in large quantities every second, mainly through social networking platforms such as Twitter and Facebook. Taking into account the fact that social data is highly dynamic and unstructured, we understand that this type of integration introduces additional technological and organisational requirements. It should be also noted that we do not expect permanent links to be established between government and social data in this type of integration. Nevertheless, the appropriate mechanisms to allow and facilitate this type of integration should be established.

The additional complexity related to this stage could be better described by a real-world case. A very popular attribute of social data that enables personalisation is the location

from which a message is published online. This attribute could be the joint point for different government and social data sets. However, the format and granularity of data describing locations can vary between different data sets. For example, although Twitter adds the longitude and latitude of a point to tweets posted by mobile applications, Ordnance Survey in the UK does not provide a service for mapping a specific point to an administrative area in order a linking between these to representations of a geo-spatial object to be enabled.

The proposed approach is based on the creation of a richer semantically enabled knowledge base from microblog data by employing Named Entity Recognition (NER) to identify entities in the text and classify them into categories as well as the Linked Data paradigm to perform entity disambiguation and to enrich microblog data with domain knowledge that currently exists as structured data on the Web e.g. through DBpedia or OGD. In order to demonstrate and evaluate the pro-posed approach we first identify and empirically study casual NER approaches and tools in microblog data realm since literature lacks a relevant study

Chapter 4

Integrate Open Data

4.1 Introduction

This chapter explores Open Government Data (OGD) integration. Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data (Lenzerini, 2002). Based on OGD characteristics we explore data integration taking into account two dimensions: (a) integrate OGD inside public sector and OGD with other data on the Web, and (b) integrate record level data as opposed to integrate aggregated multi-dimensional data cubes. In this context, in this chapter we present the following:

- *Record level data* (section 4.2): In this case, an architecture that enables data integration around real-world things is presented along with a proof-of-concept implementation that provides an integrated view of OGD from different public authorities in different administration levels. Moreover, the special case of combining OGD with Social Media data is explored and a technical architecture based on linked data along with a case study are presented.

- *Aggregated multi-dimensional data* (section 4.3): In this case, a theoretical framework to formally describe how to combine two datasets is presented. This theoretical framework defines (a) binary relations that link two cubes that are compatible to integrate, and (b) operators that map from these two cubes to a new integrated one.

4.2 Record Level Data

4.2.1 Integrate Data Inside Public Sector

4.2.1.1 Architecture

As already mentioned, Data.gov.uk is the most advanced Open Government Data (OGD) initiative that provides data in Linked Data format, establishes links between different datasets and follows an indirect data provision approach dividing public administration in functional areas. This organisational and technological reality of OGD provision in the UK's Data.gov.uk manifests the need for decentralised OGD management and provision. It is also an indication of the need for provision of integrated data across public administration. In the case of the UK government, OGD is integrated around real world objects such as laws, public agencies, points of interest etc.

Therefore, in this section we present an approach for decentralised provision of OGD that will enable information integration around real-world things such as public agencies, laws and public services. According to the proposed approach public agencies publish only information that have the mandate to manage, utilising Linked Data paradigm. As already mentioned, a suitable architecture for managing Linked Data can follow different approaches, namely centralised repositories, live look-ups and total distribution in a P2P setup. The proposed architecture can be characterised as lightweight because is based on the consumption of the provided data using mashup semantic web browsers such as Sig.ma browser (Tummarello et al., 2010). However, a more advanced client could have been developed based on a federated querying mechanism.

The proposed approach is based on the creation of links between URI aliases (i.e. identifiers that refer to the same real-world thing) in a specific sector. In this case, a high-level public authority is responsible to produce and maintain URIs for the available resources in a specific functional area. This high-level authority can be therefore also considered as a Registry of Resources (RoR) since it maintains a list of available resources in the area and assigns a URI to each of them.

Hence, the RoR could be conceptually considered as a reference dataset that all other datasets should be linked to in order a connection between URI aliases to be eventually established. The RoR could be either an actual Linked Data set providing access mechanisms or just a list of officially approved URIs to be used by public agencies.

4.2.1.2 Implementation

In order to enhance clarity of the proposed approach we employ a hypothetical use case that involves real-world OGD from public agencies and schools in Greece (Kalampokis et al., 2013d). We should underline that although the public agencies and schools mentioned in the use case are real, they have not been involved in the implementation of the proposed solution.

This use case refers to the provision of information regarding Greek schools by public agencies in different administration levels. In the current situation, these public agencies provide information that are isolated, fragmented and in some cases controversial or supplementary. The goal of this prototype implementation is to show how Linked Data can be used in OGD provision in order to alleviate existing limitations.

More specifically, in the use case scenario we consider three actors:

- The Greek Ministry of Education: this maintains a relational database with information about all schools in the country and publishes it on its website. The ministry in our implementation plays the role of the Registry of Resources (RoR).
- The 2nd Local Directorate of Secondary Education of Athens (Directorate onwards): this maintains a relational database with information about all schools in its area and publishes it on its website.
- A specific school (namely Moraitis): this publishes information about itself on its website.

The steps that are followed in this prototype implementation along with the used tools can be briefly described as follows:

- Step 1. The Directorate publishes Linked Data from the relational database. This step also includes the creation of a vocabulary for describing the schools that the directorate supervises.
- Step 2. The school publishes Linked Data in HTML embedding RDFa markup in its Web site.
- Step 3. The Ministry of Education, which plays the role of the RoR publishes a list of all Greek schools as Linked Data. This ensures the definition of URIs for all schools as well as the provision of an endpoint that facilitates the identification of URI aliases.
- Step 4. URI aliases between the RoR and the two pubic agencies are identified and typed links are created among them.

By following the specific approach, the information provided by different public sources about a specific school (e.g. Moraitis school in our case) becomes linked and citizens are able to search for and get an integrated view using e.g. semantic mashup tools such as Sig.ma. In addition, due to the typed links between the disparate sources of government data one can follow these links and receive more relevant information such as other schools in the area. In the rest of this section each step is described in detail along with the technical implementations.

Step 1: The Directorate Publishes Linked Data

The Directorate currently uses a Relation Database Management System (RDBMS) to store its data and also provides a search interface through its web site. As a result, the Directorate could use one of the available tools for publishing the content of the RDBMS as Linked Data (for a complete review of such tools consider (Sahoo et al., 2009)):

- D2R Server (Bizer and Cyganiak, 2006)
- Virtuoso RDF Views Linked Data wrapper
- Triplify (Auer et al., 2009)

For our implementation we chose D2R server since it is one of the most mature relevant solutions. D2R server is an HTTP server that can be used to provide a Linked Data

view, an HTML view and a SPARQL protocol endpoint over data stored in a relational database. D2R server uses the D2RQ language in order to map database content to RDF by a declarative mapping, which specifies how resources are identified and how property values are generated from database content.

As already mentioned, the first step in the Linked Data publishing process is the development of a domain model either by creating a new local schema (in RDF Schema, OWL, etc.) or by reusing existing, widespread vocabularies with the emphasis on the latter. D2RQ mapping language relates tables of the relational database to RDF resources and table columns to RDF properties in order to create the vocabulary for annotating RDF data. However, in some cases this simple rule is not effective and thus customisation of the initial mapping file is required.

This is also the case in the RDBMS of the Directorate. Figure 4.1 depicts the schema of Directorate's database as well the Linked Data schema created. As the figure indicates from the four tables of the relational database only three Classes were created i.e. PublicEntity, GeoArea and PublicEntityType. This is because the SupervisedEntities table is a junction table and thus it was utilised for the creation of the supervisedBy property. The specific table contains information about the fact that a particular public agency can supervise another public agency.

In addition, the basic Classes of the Linked Data schema have been linked to three well-known vocabularies i.e. FOAF, GeoNames and SKOS. More specifically the following statements have been made:

- The PublicEntity class is subClassOf the Organisation class defined by FOAF.
- The GeoArea class is subClassOf the Feature class defined by GeoNames.
- The PublicEntityType class is subClassOf the Concept class defined by SKOS.

In Figure 4.2 a piece of code is presented that describes how data contained in the PublicEntity and SupervisedEntities tables is mapped to RDF data, which is presented as a graph. For the shake of simplicity only a limited number of attributes and tuples are depicted. In particular, Figure 4.2 includes only the id, name and e-mail attributes in the PublicEntity table. We should also note that we use the "vocab" namespace

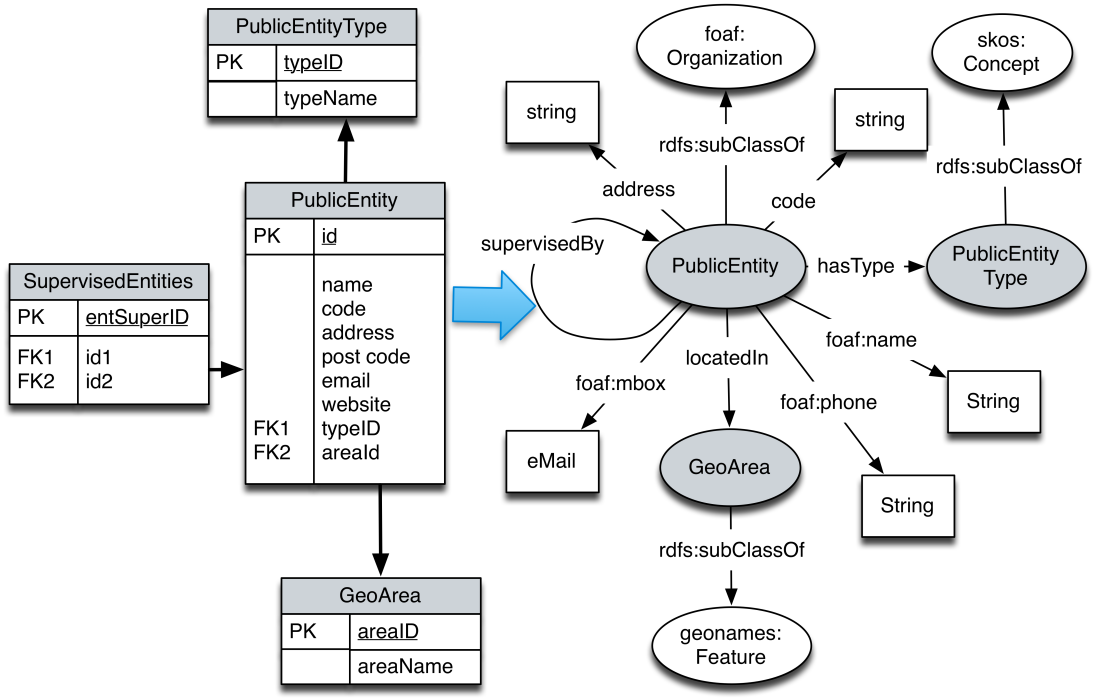


FIGURE 4.1: The mapping between Relational Schema and the Linked Data vocabulary (Kalampokis et al., 2013d)

for representing the new terms created by the Directorate to describe concepts and properties in the domain.

The main points of the process presented in Figure 4.2 can be summarised as follows:

- From each tuple in the PublicEntity table a PublicEntity resource is created. For example, the tuple described by id=1 created the resource PublicEntity/1 which belongs to the Public Entity Class. The URIs are created utilising the id attribute of the PublicEntity table.
- From each PublicEntity resource a number of properties are created based on the attributes of the PulicEntity table. For example, from the first tuple a triple is created having as subject the resource PublicEntity/1, as predicate the foaf:name property and as object the string "Attiki Prefectural Directorate of Education".
- From each tuple in the SupervisedEntities table a typed connection is created between two PublicEntity resources in the graph. For example, the first tuple in the table creates a link between the resources PublicEntity/8 and PublicEntity/1 utilising the vocab:publicentity_belongs property.

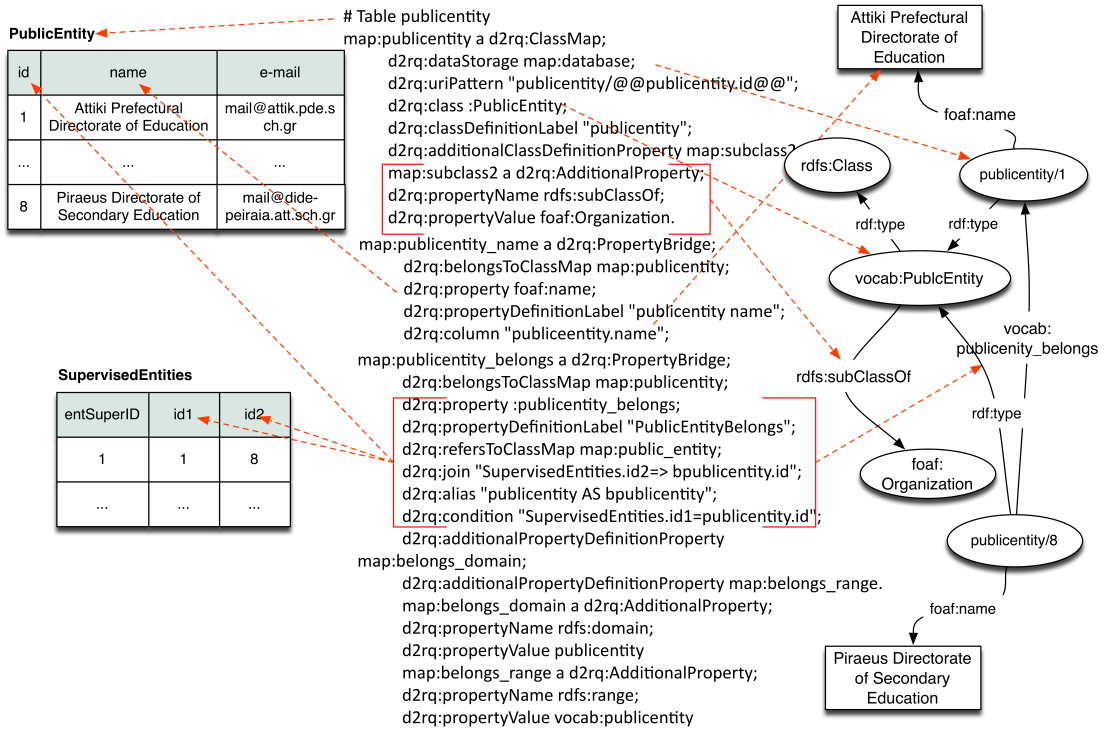


FIGURE 4.2: Using D2RQ language for creating RDF data from relational data (Kalampokis et al., 2013d)

The installation of the D2R server results in the creation of a linked data API as well as a search interface. By using the former one can refer to the non-information resource "Attiki Prefectrural Directorate of Education" using the following URI: `http://195.251.218.37:2020/resource/publicentity/1`. In addition, one can receive an RDF/XML representation describing the same resource using the following URI: `http://195.251.218.37:2020/data/publicentity/1`. Finally, an HTML representation of the same resource can be received using the following URI: `http://195.251.218.37:2020/page/publicentity/1`. In Figure 4.3, the HTML representation of the specific resources is presented. This interface enables users to move to the representation of different resources using the vocab:publicentity_belongs property.

As regards the search interface, a SPARQL endpoint is provided in the following URI: `http://195.251.218.37:2020/snorql`. This endpoint enables users to execute more complex queries such as "What is the phone and the e-mail address of the public agencies that are supervised by the First Directorate of Secondary Education in Athens?". The answer to this question could be given by the following SPARQL query:

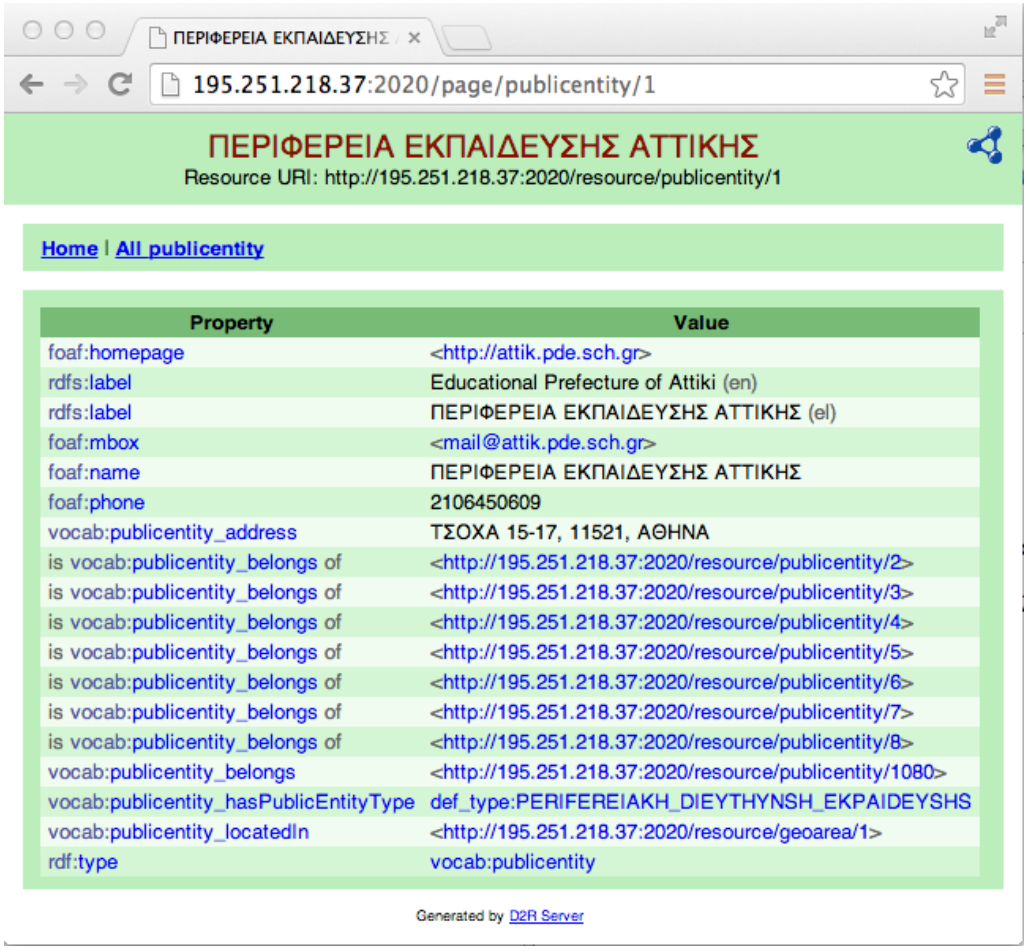


FIGURE 4.3: A description of the Educational Prefecture of Attiki using D2R server

```
SELECT DISTINCT ?mail ?phone WHERE {
?a  vocab:publicentity_code '201' .
?s  vocab:publicentity_belongs ?a ;
     foaf:mbox ?mail ;
     foaf:phone ?phone .
}
```

Step 2: The School Publishes Linked Data

Moraitis School maintains a web page where it provides information about the school. The page has been created using HTML and thus the most efficient way to add semantics into it is the use of RDFa. In particular, RDFa embeds markup data within a Web document in order to make it understandable for machines as well as people.

The basic rules of RDFa markup are the following:

- If the object of the statement takes a literal string as its value, this literal string will be the value of content attribute, the subject will be the value of about attribute and the predicate will be given by the value of property attribute.
- If the object of the statement takes a resource as its value, the resource will be identified by the value of href attribute, the subject will be the value of about attribute and the predicate will be given by the value of rel statement.

A part of the HTML code used by Moraitis school before the incorporation of RDFa markup is presented below.

```
<!-- other HTML code -->
<div>
     <p>Moraitis School is located in Psychico, Athens</p>
     <img src=http://www.moraitis-school.com/school/images/Sxoleio02.jpg/>
     Email: <a href=‘‘mailto:info@moraitis-school.com’’>info@moraitis-school.com</a>
</div>
<!-- other HTML code -->
```

After inline RDFa to markup, the page content will look like the following piece of code. We should note that the Web page presented to a human user from these two pieces of code is identical.

```
<!-- other HTML code -->
<div xmlns:foaf=‘‘http://xmlns.com/foaf/0.1/’’
typeof=‘‘foaf:Organization’’ about=‘‘http://www.moraitis-school.com/#moraitis’’>
<span property=‘‘foaf:name’’>Moraitis School </span>is located in
<span  rel=‘‘foaf:based_near’’ href=‘‘http://dbpedia.org/resources/Psychico/’’/>
Psychico, Athens</span>

<span rel=‘‘foaf:depiction’’>  <img src=http://www.moraitis-school.com/school/
images/Sxoleio02.jpg /> </span>

Email: <a rel=‘‘foaf:mbox’’ href=‘‘mailto:info@moraitis-school.com’’>
info@moraitis-school.com</a>
```

```
</div>
<!-- other HTML code -->
```

The school uses FOAF vocabulary to describe information about itself. Although the main aim of FOAF is to describe people, it is used also for organisations since both foaf:Person and foaf:Organization classes are sub-classes of foaf:Agent class.

Step 3: Creating the Registry of Resources

The final step of the process is the creation of the Registry of Resources (RoR), which we assume that will be implemented by the Greek Ministry of Education. The RoR publishes a list of all public agencies supervised by the Ministry of Education using Linked Data format. As a result, the RoR assigns a dereferenceable URI to every public agency and thus to every school in Greece.

We assume that the Ministry maintains a database containing the name and the official identification number issued by the government for every school in the country. In this case, the D2R server is installed in order to publish the specific information as Linked Data. The vocabulary describing this installation is indicated in the rest of the article by the namespace "ror". The results of the D2R installation in RoR's server are the provision of a Linked Data API as well as a SPARQL endpoint.

It should be noted that the URIs created by the RoR are different from the ones created by the Directorate because of the unique identifier used. In particular, the Directorate used the value of the id attribute of the PublicEntity table (i.e. the URIs follow the format /publicentity/{id}), while the RoR uses the official identification number issued by the government (i.e. the URIs follow the format /school/{code}).

Step 4: Linking the Resources

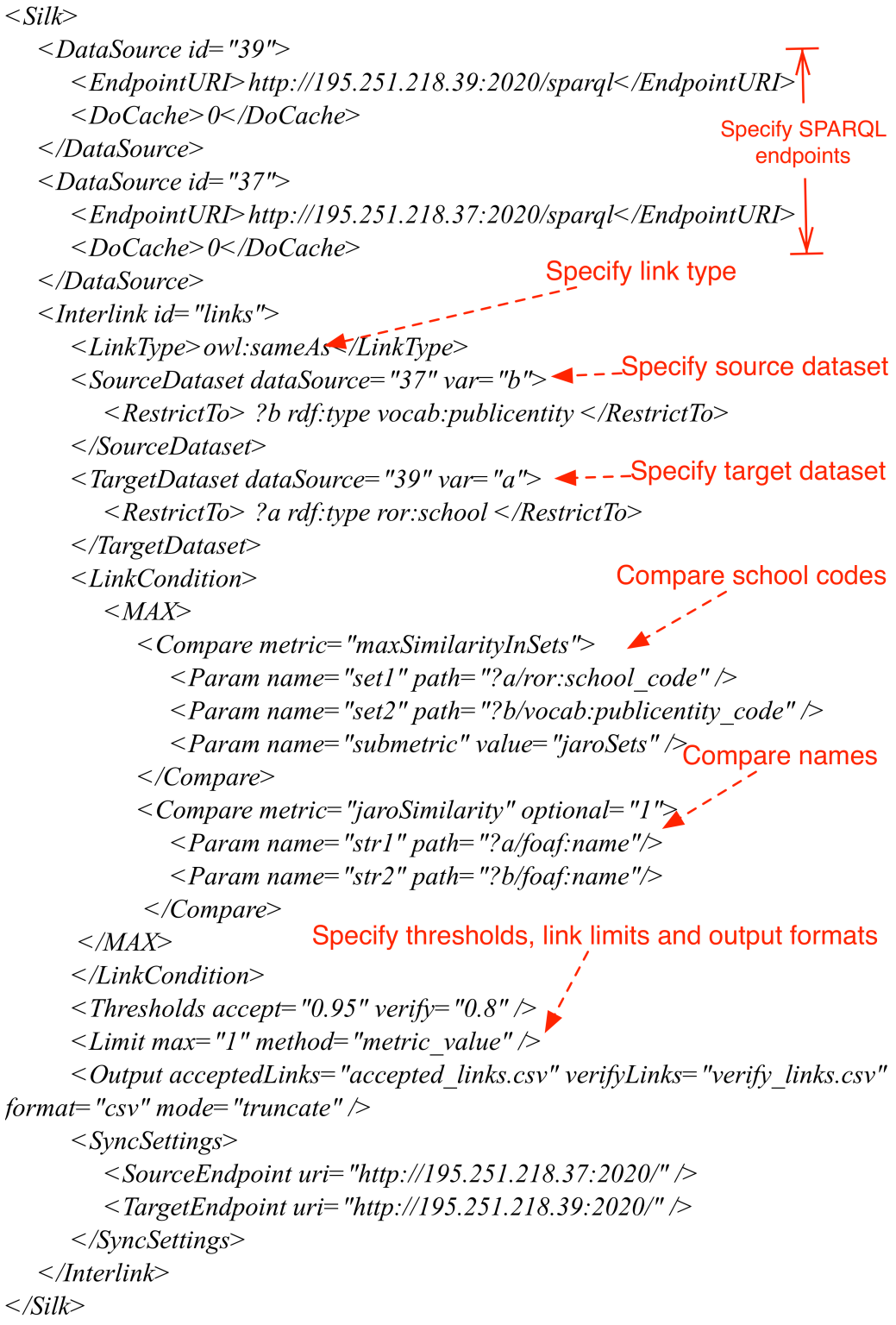
The final step aims at identifying URI aliases between the RoR and the pubic agencies and creating typed links between them. The most prevalent way of dealing with URI aliases is to use the owl:sameAs predicate to link between them. An owl:sameAs statement indicates that two URI references actually refer to the same thing. For doing this mapping between the different data sets we employ the Silk framework (Volz et al., 2009). The Silk framework uses the declarative Silk–Link Specification Language (Silk–LSL) so that data publishers to be able to specify which types of RDF links should be

discovered between data sources as well as which conditions data items must fulfill in order to be interlinked. These link conditions can apply different similarity metrics to multiple properties of an entity or related entities, which are addressed using a path-based selector language. The resulting similarity scores can be weighted and combined using various similarity aggregation functions. Silk is accessing the two data sources that we want to link through their SPARQL endpoints. In the rest of this sub-section we describe how the Directorate and the School are linked to the RoR.

First the Directorate aims to link its data to data provided by the RoR. In particular, the Directorate aims to identify and link URI aliases between these two datasets. In Figure 4.4 we present the code used for creating the links in the Link Specification Language of Silk.

The main points of the linking process as described in Figure 4.4 are the following:

- Specification of the SPARQL endpoints of the two data sources i.e. the Directorate and the RoR.
- Specification of the link type that will be used to connect data from the two data sources. In our case we use the owl:sameAs predicate.
- Specification of the source and target data sets as well of the resources that will be used for the matching. Since we only want to match schools, we restrict the sets of examined resources to instances of the classes vocab:PublicEntity and ror:School by supplying SPARQL conditions within the <RestrictTo >directives.
- Definition of how similarity metrics are combined in order to calculate a total similarity value for an entity pair. Each metric in Silk evaluates to a similarity value between 0 or 1, with higher values indicating a greater similarity. In our use case we used the maxSimilarityInSet metric, which returns the highest encountered similarity of comparing a single item to all items in a set, to match school codes and the jaroSimilarity metric, which is a string similarity based on Jaro distance metric, to match the school names. The <MAX >tag enables the combination of the two similarity metrics by choosing the highest value.
- Specification of the threshold, link limits and output formats. As regards the threshold we specify that resource paris with a similarity score above 0.95 are to




```
<Silk>
    <DataSource id="39">
        <EndpointURI>http://195.251.218.39:2020/sparql</EndpointURI>
        <DoCache>0</DoCache>
    </DataSource>
    <DataSource id="37">
        <EndpointURI>http://195.251.218.37:2020/sparql</EndpointURI>
        <DoCache>0</DoCache>
    </DataSource>
    <Interlink id="links">
        <LinkType>owl:sameAs</LinkType>
        <SourceDataset dataSource="37" var="b">
            <RestrictTo> ?b rdf:type vocab:publicentity </RestrictTo>
        </SourceDataset>
        <TargetDataset dataSource="39" var="a">
            <RestrictTo> ?a rdf:type ror:school </RestrictTo>
        </TargetDataset>
        <LinkCondition>
            <MAX>
                <Compare metric="maxSimilarityInSets">
                    <Param name="set1" path="?a/ror:school_code" />
                    <Param name="set2" path="?b/vocab:publicentity_code" />
                    <Param name="submetric" value="jaroSets" />
                </Compare>
                <Compare metric="jaroSimilarity" optional="1">
                    <Param name="str1" path="?a/foaf:name"/>
                    <Param name="str2" path="?b/foaf:name"/>
                </Compare>
            </MAX>
        </LinkCondition>
        <Thresholds accept="0.95" verify="0.8" />
        <Limit max="1" method="metric_value" />
        <Output acceptedLinks="accepted_links.csv" verifyLinks="verify_links.csv"
format="csv" mode="truncate" />
        <SyncSettings>
            <SourceEndpoint uri="http://195.251.218.37:2020/" />
            <TargetEndpoint uri="http://195.251.218.39:2020/" />
        </SyncSettings>
    </Interlink>
</Silk>
```

FIGURE 4.4: The SILK-LSL code used for creating links

be interlinked, whereas pairs between 0.8 and 0.95 should be written to a separate output file and be reviewed. The link limit is used to limit the number of outgoing links from a particular entity within the source data set. In particular, only one outgoing owl:sameAs link is permitted. Finally, the output format is specified.

The result of this step is a CSV file containing the links above the accepted threshold. Because we use the unique identification number, the accepted matches are more than 98 percentage of the total number of schools. Finally we import the CSV file in the relational database of the RoR. As a result the dataset provided by the RoR and the dataset provided by the local directorate are connected through owl:sameAs links between URI aliases.

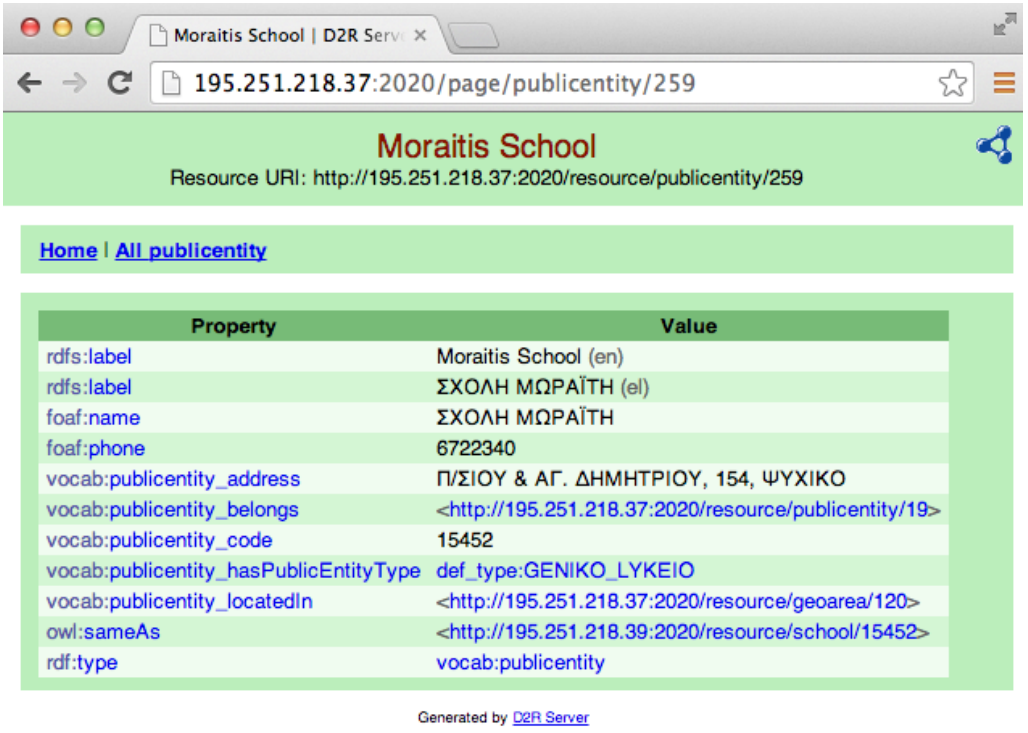


FIGURE 4.5: The representation of Moraitis School after the insertion of the owl:sameAs link

In Figure 4.5 the representation of Moraitis School from the Directorate is presented. Also the incorporation of a link to the RoR through the owl:sameAs predicate is emphasised. As regards the School, the linking of the resource described the Web page and the respective resource in the RoR can be done manually by adding the following piece of code in the HMTL page presented in Step 2:

```
<span rel=owl:sameAs resource=http://195.251.218.39:2020/resource/school/15452 />
```

More specifically, this RDFa markup indicates that the resource that describes is owl:sameAs the entity Moraitis School described in the RoR (ror:school/15452).

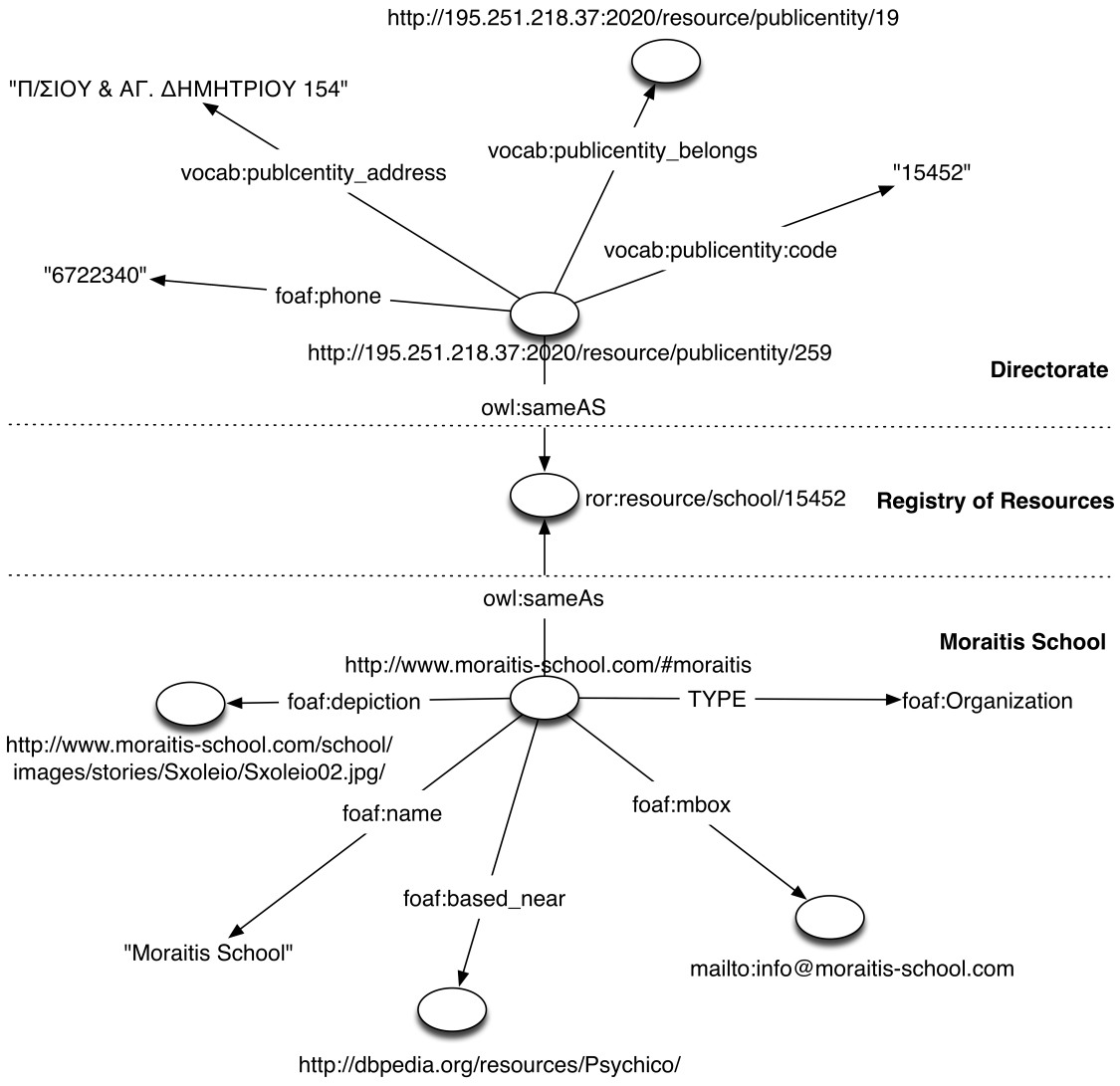


FIGURE 4.6: The linked data graph

4.2.1.3 Integrated view of OGD

The final outcome of these steps as regards the School of Moraitis related data is the creation of the graph depicted in Figure 4.6. It presents the linking of OGD about the specific school provided by two distributed sources i.e. the 2nd Local Directorate of

Secondary Education of Athens and the school itself. The ror:school/15452 representation of the school is provided by the RoR and it is the glue between the representations http://195.251.218.37:2020/resource/publicentity/259 and http://www.moraitis-school.com/#Moraitis since these two are linked to the first one by two owl:sameAs links. By following the specific approach, the data provided by different public sources about Moraitis School is now linked and the data consumer can search for and get the integrated view by using semantic mashup tools such as Sig.ma. In addition, due to the typed links between the disparate sources of government data one can follow these links and receive more relevant information such as other schools in the area.

4.2.2 Integrate OGD with Social Media Data

In this sub-section we present the case of integrating OGD with social media (SM) data. We consider this type of integration as record-level data integration because it takes place based on real world entities and their characteristics. Regarding SM data the integration can be based either on their metadata or the actual content (e.g. the text of a tweet). At the latter case, Named Entity Recognition (NER) analysis is required to extract from the text the entities that will be used for the integration.

In both cases the integration of SM data with open data aims at creating a richer knowledge base that will enable more valuable analyses. In the following sub-sections this idea is further illustrated.

4.2.2.1 Integration Based on Metadata

SM data is streamed in large quantities every second, creating significant information overload for the users interested in making sense of the information related to a specific context. This is particularly true in decision-making where decision makers want to listen to people that are expressed about a specific topic of interest or/and are affected by a particular decision, and not to the whole population. So, after its collection, SM data should go through a filtering stage in order to be narrowed based on some criteria.

Existing approaches in the literature use keyword search or hashtag search in order to alleviate the problem of information overload. However, this type of search can only support the selection of SM data related to a specific topic of interest, e.g. immigration,

or to a specific event, e.g. publication of a new draft law. In our approach, we want to enhance such solutions with capabilities that will enable the filtering of social data based on the target group i.e. people affected by a particular decision. This could include for example, the identification of data created by female users above the age of 18 or the identification of data created by citizens that live in areas characterised by high crime levels.

To this end, we propose that we should enrich SM data with OGD (Kalampokis et al., 2011a). Characteristics of target groups such as age group, gender and area of residence could be linked to variables included in government datasets that provide objective facts related to these characteristics.

In order to make our point clear we now describe a real-world scenario. According to this scenario the government of the UK announces to citizens a draft law on public budget cutting in police forces. Before the enactment of the particular draft the government wants to know what citizens think about the specific action. Moreover, the government is particularly interested in the opinion of residents of areas presenting crime level above average.

According to our approach, subjective data will be collected from Twitter before and after the announcement of the draft law. In order to identify only those tweets that are posted by residents of areas with crime level above average we will aggregate data from Data.gov.uk that provides crime levels and statistics in neighbourhood areas in the 43 English and Wales' police forces through a RESTful API and data from Twitter. By linking the location attribute of tweets to the "crime area" attribute of the Data.gov.uk dataset we can filter the collected tweets and identify tweets posted by residents of areas with high crime level. Figure 4.7 depicts the linking of the two datasets using as a "joint point" the particular location i.e. Leicestershire.

4.2.2.2 Integration based on Entities Extracted from Text

The informal and noisy nature of SM result in spelling mistakes and user made acronyms and abbreviations that is not possible to be identified by keyword search. For example, the Liberal Democrat political party in the UK can be mentioned as liberal democrats, libdem, libdems, ld, etc. Moreover, some complex phenomena such as elections are

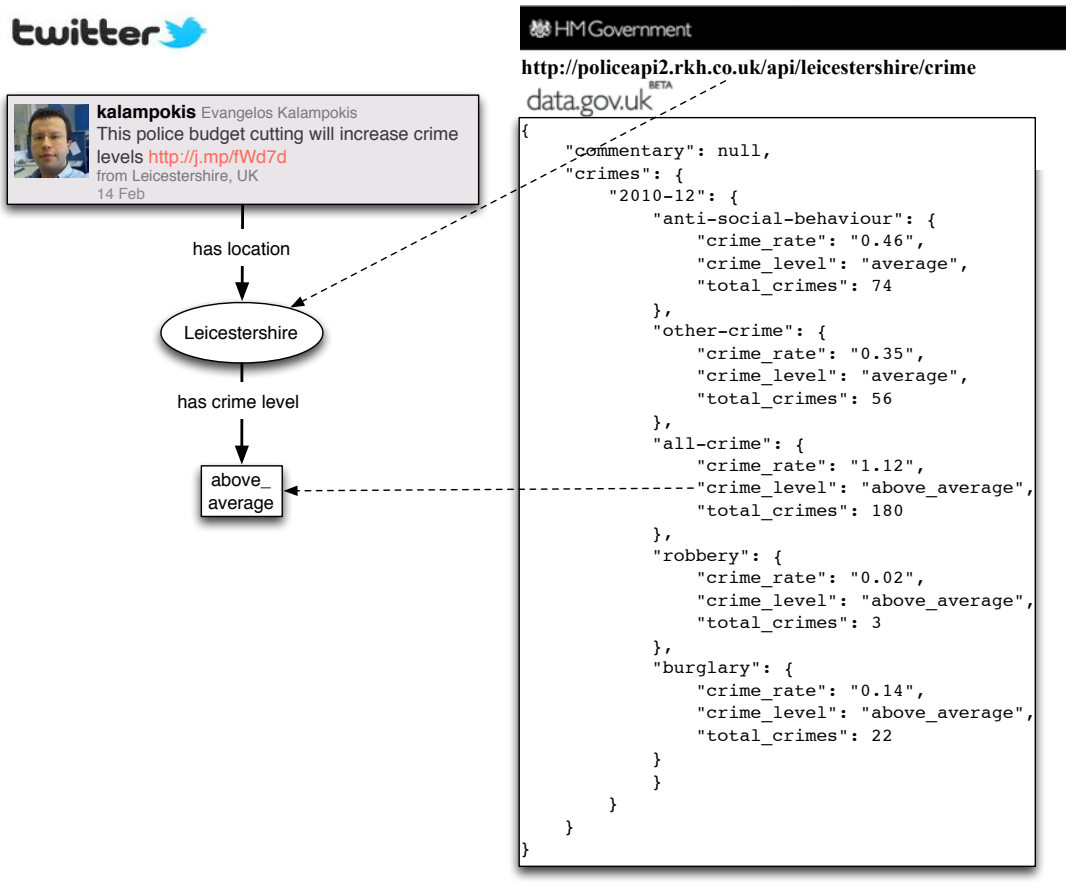


FIGURE 4.7: Filtering tweets based on the crime level of the location (Kalampokis et al., 2011a)

characterised by many and diverse participating entities. For example, in order to sense public opinion regarding a particular political party, one should identify tweets that mention not only the political party but also the leader of the party as well as the candidates of the party in different geographical areas or even former leaders or colloquial names. In this case all the different keywords that describe this domain knowledge should be identified and used. However, this activity may introduce subjectivity in the analysis process. Finally, keyword search is not context aware and thus provides irrelevant results, e.g. in the context of the UK elections the word 'Gordon' can refer to both the homonymous UK parliament constituency in Scotland and the first name of Gordon Brown, who was the leader of the Labour Party from 2007 to 2010.

In this sub-section we present an approach to create a rich semantically enabled knowledge base from Twitter data that will enable more efficient identification of relevant data in the context of a phenomenon under consideration. Following the UK election use case the proposed approach will enable answering questions such as "Which posts are related

to the conservative party" instead of "Which posts include the strings 'conservative', 'cameron', 'tories' etc.". As a result, the analysis process will be released from subjective decisions, such as the identification of relevant keywords, which could influence the whole process and thus the outcome.

The proposed approach is based on the creation of a richer semantically enabled knowledge base from tweets by employing Named Entity Recognition (NER) to identify entities in the text and classify them into categories, as well as the Linked Data paradigm to perform entity disambiguation and to enrich SM data with domain knowledge that currently exists as structured data on the Web e.g. through DBpedia or OGD. As a result, our approach goes through two main steps:

- Identification and classification of the entities that are mentioned in microblog data.
- Identification of URI aliases of these entities on the Web and establish links among them.

In order to make the proposed approach more clear we again employ the UK election of 2010 use case. Figure 4.8 depicts four tweets that were published before the election. Four named entities related to the election are extracted from the text i.e. Cameron, George Osborn, Tory and Conservatives. These entities are linked to the same entities that are described in DBpedia (i.e. dbpedia:David_Cameron, dbpedia:George_Osborne, dbpedia:Tory and dbpedia:Conservative_Party_UK respectively) by establishing owl:sameAs links. As Figure 4.8 indicates the latter entities are connected between them through RDF links. In particular DBpedia suggests that David Cameron and George Osborne belong to the Conservative party and that the Conservative Party is also known as Tory party. Following this approach one can identify all the tweets that refer to the Conservative party even if the specific keyword is not directly mentioned in their text.

In Figure 4.9 an architecture that enables the realisation of the proposed approach is depicted. The architecture comprises a number of components. The Microblog SM Data Mining component collects SM data through different APIs. The collected data contains all the attached metadata (e.g. text, author, creation date, source, geo-location information, etc.) along with the actual text. The Extraction component comprises two modules i.e. the Metadata Extractor and the NER Extractor. The former extracts the

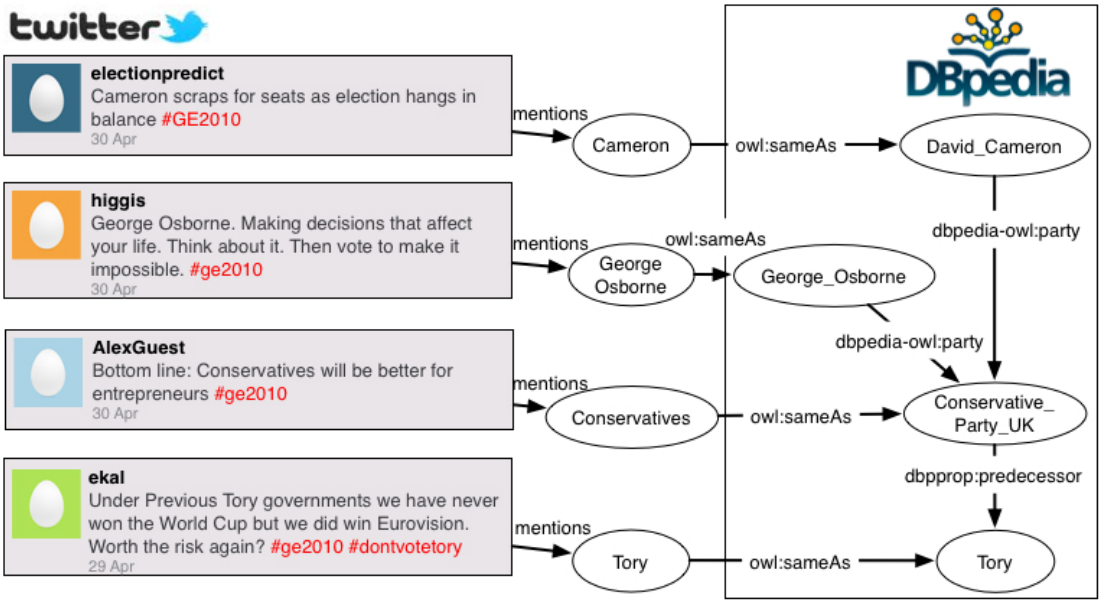


FIGURE 4.8: Extracting entities form tweets and linking them to URI aliases in DBpedia

metadata while the latter the entities in the text by employing NER techniques. The RDFiser component transforms the data into RDF and annotates the data with well-known vocabularies and ontologies. To this end we employ widely used vocabularies such as SIOC. For example, every post is an instance of sioc:MicroblogPost and the author is an instance of sioc:UserAccount. The original text is described by the sioc:content property and the entities extracted from the text by the sioc:topic property. We also differentiate between the entities of different types (i.e. person, organisation, location) and we assign them different classes. The output of this component is RDF triples, which are stored in the triple store. The Interlinking component identifies URI aliases on the Web for the extracted entities and links them through owl:sameAs links. The aim of this component is to perform a) entity disambiguation i.e. to specify the extracted entities that refer to the same real-world entity and b) data enrichment by enabling the use of existing linked data on the Web. In addition, because of the informal and noisy nature and short length of microblog posts we adopt a two-phased interlinking process. Initially based on a strict matching algorithm such as Levenshtein distance or Jaro-Winkler distance we interlink the extracted entities with entities on the Web that are of the same named type as the extracted one. Thereafter with a looser matching algorithm such as Boyer Moore algorithm we interlink the extracted entities that were

missed during the first phase with the same entities that we identified on the Web in the first phase. Each entity in the data set carries two features: a) its class (person, organisation or location) and a) its literal value. The features of every unlinked entity are compared with the features of every interlinked entity. If there is a suitable match then the unlinked entity is linked to the similar entity by owl:sameAs link.

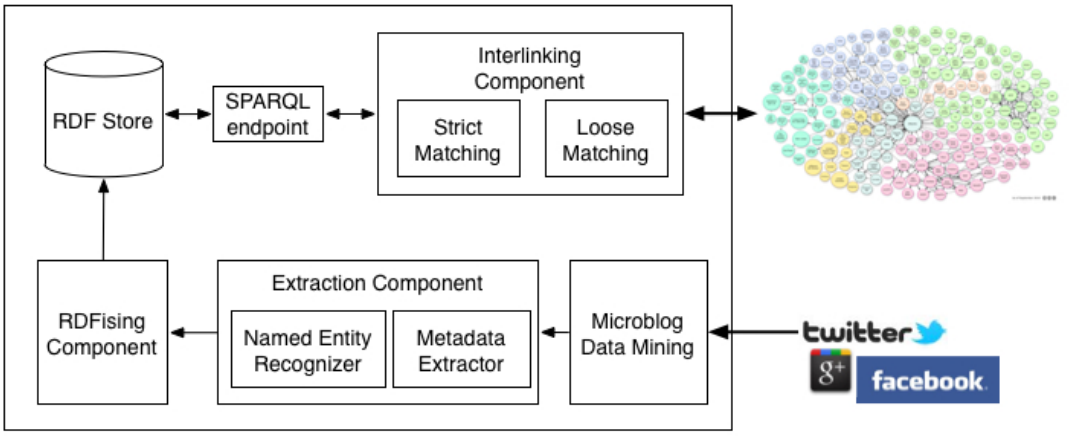


FIGURE 4.9: Architecture showing the processing of microblog data into Linked Data

We implement the proposed architecture and analyse 10.591 tweets that were randomly sampled from a UK election data set. In order to create this dataset we collected 60.000 unique tweets from Twitter using #ge2010, #ukelection, #election2010 and #ge10 hashtags from April 29, 2010 to May 6, 2010.

The evaluation task is to identify tweets that are related to the three main parties of the UK election (i.e. Conservative Party, Labour Party and Liberal Democrats). We compare the number of the identified tweets through this process with the number of tweets that would be identified through keyword search.

The implementation of the approach includes a Conditional Random Field sequence classifier provided by Stanford NER in order to put into practice the Named Entity Recogniser of our architecture (In the Appendix F we present an evaluation of 8 NER tools that justifies the selection of the Stanford NER component). The classifier is specially trained for the specific data set that we work on and achieves 85.89 F1 score. Currently this component extracts 3 NET: Person, Location and Organisation. The Interlinking Component aims at identifying URI aliases with DBpedia. To this end, we employ Silk Framework (Volz et al., 2009) to realise the first Strict-matching phase of our

architecture. The unlinked entities in this first phase of interlinking go through a second phase that according to the architecture is characterised as Loose-matching phase. This phase implements Boyer-Moore algorithm and the unlinked entities are linked to the similar DBpedia entity (of the same class) found on the first phase. Finally, URI aliases between the two data sets are connected and the data set is updated using SPARQL Update. For the sake of performance (faster SPARQL queries) we establish RDF links in the local RDF data store using DBpedia properties. We currently use a Jena SDB backed by MySQL database to store the RDF data set. Joseki server is used to provide a SPARQL endpoint to this SDB store.

The result of the implementation is a semantically enabled knowledge base that could be accessed through a SPARQL endpoint. The total number of distinct entities for each NET in the produced dataset are: 30 Organisation (e.g. dbpedia:Liberal_Democrats, dbpedia:Sky_News, dbpedia:BBC) 286 Person (e.g. dbpedia:Gordon_Brown, dbpedia: David_Cameron) and 230 Location (e.g. dbpedia:Edinburgh, dbpedia:Islington, dbpedia:Bradford).

Initially we searched the raw data set by using the following keyword: "conservative" for the Conservative party, "labour" for the Labour party and "liberal" and "democrat" for Liberal Democrats. Thereafter we queried the produced RDF data set based on a SPARQL query. For example, for Liberal Democrats the following query was used, where twit:microblogOrgEntity represents the Organisation named entity type:

```
SELECT DISTINCT ?tweet
WHERE {
{
?tweet sioc:topic ?org.
?org rdf:type twit:microblogOrgEntity.
?org owl:sameAs <http://dbpedia.org/resource/Liberal_Democrats>
}
UNION {
?tweet sioc:topic ?otherEntity.
?otherEntity owl:sameAs ?dbpEntity.
?dbpEntity ?rel <http://dbpedia.org/resource/Liberal_Democrats>
}
```

```
}
```

The first part of the query finds all those mentions of Liberal Democrats that directly mention the entity Liberal Democrats. The second part of the query finds those mentions which indirectly mention the entity Liberal Democrats, for e.g. a tweet that mentions Nick Clegg is supposed to refer to Liberal Democrats indirectly because Nick Clegg is the leader of Liberal Democrats.

Figure 4.10 depicts the evaluation results for the three parties. The first column presents the identified tweets with keyword search. The second and third columns present the tweets identified through entity disambiguation after the application of Strict-matching and Loose-matching phase of the architecture respectively. The last column depicts the final results obtained from our approach after also incorporating DBpedia's domain knowledge. As Figure 4.10 suggests the number of tweets that were identified with the proposed approach was doubled in the case of the Labour party and quadrupled in the case of both the Conservative party and Liberal Democrats. These results indicate that keyword search would have provided misleading data to the next steps of the prediction analysis phase.

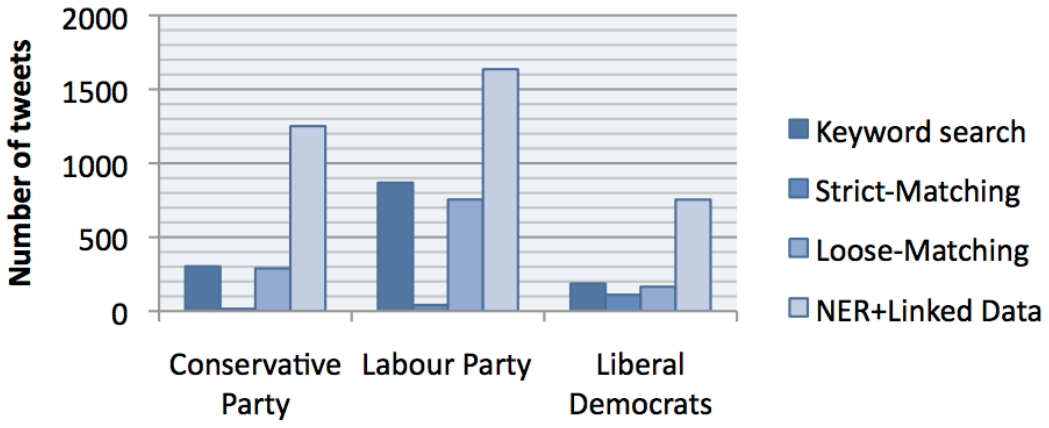


FIGURE 4.10: The evaluation results

In addition, the evaluation results indicate that the results after the Strict-matching phase are quite poor. This is due to the fact that not many entities in the local data set are linked to DBpedia entities when Silk is applied. This happens because Silk uses Approximate String Matching Algorithms such as Levenshtein distance and Jaro-Winkler distance. In such algorithms operations such insertion, deletion and substitution

of characters in strings are carried out to convert a string into an exact match. The less the number of these operations, closer the match is. For example insertion of u in labor, deletion of e in laboure and substitution of o with a at 1st index in lobor, all will lead to the exact match labour. However, if the string is lab, it will be considerably much difficult to match it with labour. Therefore, Silk was not able to match the entities which were very different in length, for e.g. lab and labour, liberal and liberal democrats, david and david Cameron, etc. The Loose-matching column clearly indicates a significant improvement in the results. The Loose-Matching phase which uses an Exact String Matching Algorithm is now able to link the entities like lab with labour; lib or liberal or dem or democrats with liberal democrats and many such other entities which Silk failed to link. However, we still miss out on interlinking the entities like liberals or libdem with liberal democrats, which require more complex algorithms to be matched.

As regards Liberal Democrats after the Strict-matching phase we get 111 tweets that refer to the entity Liberal_Democrats, which has liberal democrats as the string value for the rdfs:label property. After the second interlinking phase we get 54 additional tweets of the entity Liberal Democrats. These mentions include entities like Lib, LIB, DEM and Democrats. The number of tweets increases to 754 after DBpedia's domain knowledge is also incorporated in our data set. The additional tweets are related to Liberal Democrats through some property. For example, Person entities such as Julian_Huppert, Nick_Clegg, Vince_Cable, Sarah_Teather and Charles_Kennedy are related to Liberal Democrats by dbpedia-owl:party property and the Organization entity Labour_Party(UK) by dbpedia-owl:mergedIntoParty property.

This analysis of SM and open data integration will be further used in a data analytics case in Chapter 5.

4.3 Mutli-Dimensional Data Cubes

In this section, we introduce a theoretical framework that formally describes the integration of OGD that are structured as aggregated multi-dimensional data cubes (data cubes onwards). In particular, we introduce the concept of "cube expansion", which defines how a cube could *expand* another cube through their integration. This is similar to a left outer join in SQL. Thereafter, we elaborate on the use of the linked data paradigm as a

facilitator to cubes integration. We define a conceptual process model that exploits the theoretical framework and introduces cubes integration as a vital phase into linked data cubes lifecycle. Finally, we identify and present challenges related to the exploitation of linked data paradigm for cubes integration.

Although cubes integration has being studied in data-warehouses literature for more than a decade, OGD have introduced new requirements in the area. Typically, an organisation had a collection of measures that are important to track for an area and based on these a data-warehouse were being developed. In open data realm, however, data providers make available for reuse in an ad-hoc manner multiple datasets that may contain parts of a bigger cube with multiple measures, dimensions, and hierarchies. On the other hand, however, users may need data that require the integration of these datasets or even the data cubes that can be created by integrating the datasets. As a result, cube integration has to be studied under this new perspective.

One of the earliest definitions of cubes integration was introduced by Agrawal et al. (1997). In this work the join operation relates two cubes and it is based on a number of dimensions which are called the *join dimensions*. In the most generic case, the two cubes need to have at least one join dimension. For example, Figure 4.11 illustrates cube C joining with cube C1 on dimension D1. Dimension D1 of the resulting cube has only two values. Moreover, a function is associated with the join that divides the element value from cube C by the element value from C1. The interesting trait here is that the values from C1 are applied to all member of the non-joinable dimension of C. In the case that there are no join dimensions this type of join is considered to be a Cartesian product between the cubes. Another special case appears if all dimensions of one cube are joined with some of the dimensions the other cube. This is called an associate join.

Datta and Thomas (1999) defined two operators that enable integration of two cubes. In particular, they defined the *join* operator as the result of the Cartesian product operation to two cubes having on or more dimensions in common and having identical mappings from the common dimensions to the respective attribute (level) sets of these dimensions. The result of this operation is a superset of the desired information. A similar description of the Cartesian product is provided by Cabibbo and Torlone (1998). Datta and Thomas (1999) also defined the *union* operator that finds the union of two cubes that have the

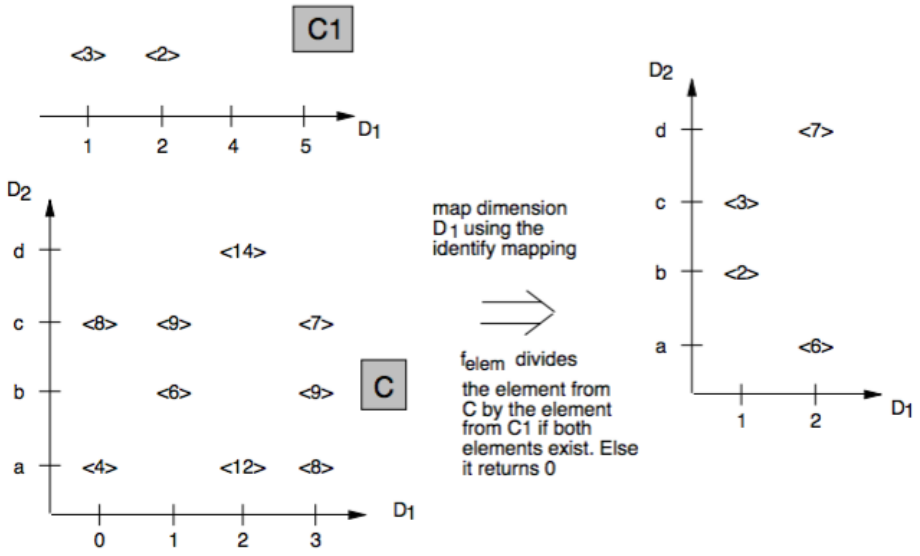


FIGURE 4.11: Joining two cubes (adopted from Agrawal et al. (1997))

same dimensions and measures but contain the superset of observations. In both cases the operators require as input compatible cubes.

Pedersen et al. (2001) also defines union and join operators between two cubes. Given two multidimensional objects with common schemas (i.e., the same set of dimensions) the union is defined as the set of union of the facts along with the union of the levels and values of the dimensions. They also provide a description of a generic type of join, called "identity-based join".

In Calvanese et al. (2001) information integration is described in DW as data passes from different internal to the organisation sources to the warehouse. The *drill-across* operation was introduced to address the case where two or more fact tables that share dimensions are combined into a single report Gómez et al. (2012); Kimball and Ross (2011). introduced *drill-across* operation define *drill-across* operation that enables combining information contained in two cubes. This operation requires that

In most of these theoretical frameworks cubes integration was only presented as part of a generic framework aiming at conceptualise cubes and thus they do not describe in detail cubes integration. Moreover, these works were purely theoretical and did not mean to be a basis for implementation.

4.3.1 A Theoretical Framework for Cubes Integration

Let X, Y and Z be three sets of cubes where $x \in X$ is a cube to expand, $y \in Y$ is a cube that is merged with x in order to expand the latter, and $z \in Z$ is the expanded form of cube x that is produced as a result of this process. In this case we say that we *"expand x by y"*.

We consider that a cube is defined as (M, D) where $M = \{m_1, m_2, \ldots, m_{\mathrm{k}}\}$ is a set of k measures and $D = \{d_1, d_2, \ldots, d_{\mathrm{n}}\}$ is a set of n dimensions. A dimension $d_i \in D$ comprises a set of objects $O_i = \{o_{i,1}, o_{i,2}, \ldots, o_{i,m}\}$ where m denotes the size of O_i that may vary depending on the dimension.

The objects of a dimension may structure hierarchical relationships among them at different levels that define the attributes of a dimension. Therefore, a dimension d_i is composed of a set of l attributes $A_i = \{a_{\mathrm{i,1}}, a_{\mathrm{i,2}}, \ldots, a_{\mathrm{i,l}}\}$, while an attribute $a_{\mathrm{i,j}}$ comprises all the objects of the dimension at the j-th level of the dimension d_i. Therefore, the set of objects of a dimension is the union of the sets of objects of all the attributes in the dimension $O_i = \bigcup_j a_{i,j}$.

The ordered set of attributes defines the hierarchy h_i of a dimension d_i. We assume that the 1st attribute of a hierarchy denotes the top attribute of a dimension.

The cube is composed of all cells $c_h = (t_{c_h}, v_{c_h})$, where $t_{c_h} \in \prod_i O_i$, meaning the generalised Cartesian product of O_i for the n dimensions of the cube, and v_{c_h} is a tuple of values for the set of measures M.

For example, let x_1 be a cube with three dimensions $D_{\mathrm{x_1}} = \{geo, time, sex\}$ and one measure $M_{\mathrm{x_1}} = \{unemployment\}$. The dimensions are characterised by the following:

- $O_{\mathrm{geo}} = \{BE, GR, IT\}$, $O_{\mathrm{time}} = \{2014, 2014-Q1, 2014-Q2, 2014-Q3, 2014-Q4\}$, and $O_{\mathrm{sex}} = \{M, F\}$
- $A_{\mathrm{geo}} = \{country\}$, $A_{\mathrm{time}} = \{year, quarter\}$, and $A_{\mathrm{sex}} = \{sex\}$

Therefore, $a_{\mathrm{time,quarter}} = \{2014-Q1, 2014-Q2,\ 2014-Q3,\ 2014-Q4\}$ and $a_{\mathrm{time,year}} = \{2014\}$.

Finally, a cell of cube x_1 could be defined as $c_1 = ((BE, 2014-Q1, M), (13))$.

We assume that a cube can be expanded by increasing the size of one of the sets that define a cube. Therefore, cube $x \in X$ can be expanded by adding one or more elements into: (a) the set of measures M_{x}, (b) the set of objects of an attribute $a_{\mathrm{i,j}}$ of a dimension $d_{\mathrm{i}} \in D_{\mathrm{x}}$, (c) the set of attributes A_{i} of a dimension $d_{\mathrm{i}} \in D_{\mathrm{x}}$, or (d) the set of dimensions D_{x}.

The prerequisite for this process is that cube y has some characteristics that allow for merging with cube x. These characteristics are constraints on the structure of the cube. We say that if cube y have these characteristics, then *"y is compatible to expand x"*.

These characteristics, however, are different in each of the four cases that can be followed to expand x. Therefore, it is possible y to be compatible to expand x in a particular way (e.g. by adding a new element in M_{x}) but not in a different one (e.g. by adding a new element in A_{i} of a dimension d_{i}). Moreover, the characteristics of cube z, also, depend on the case that will be followed to expand x.

Hence, there is a need to formally describe (a) the characteristics of cubes $x \in X$ and $y \in Y$, so that y to be compatible to expand x, and (b) the characteristics of cube $z \in Z$ in relation to x and y in each of the four cases. In the rest of this section we describe these characteristics by formally defining, respectively, (a) binary relations that link cubes x and y, and (b) operators that map from (x, y) to z.

In general, a binary relation R is an ordered triple (X, Y, T) where $T \subset X \times Y$ and $X \times Y$ is the Cartesian product of X and Y. In this case, we use the statement $R(x, y)$ where $(x, y) \in T$. On the other hand, any mapping from T to V is called an operator and we indicate $A : T \to V$.

4.3.1.1 Identifying Compatible Cubes

We define a binary relation L as an ordered triple (X, Y, G) where $(x, y) \in G$ is the set of $x \in X$ and $y \in Y$ where y is compatible to expand x. We say that *"y is compatible to expand x"* with the statement $L(x, y)$.

It does not always hold that if y is compatible to expand x then x is also compatible to expand y. In the special, however, case that this condition holds for a set (x, y) then we say that x and y are *mutual compatible*.

In this subsection we describe how the binary relation L is modified according to the different cases we could follow to expand $x \in X$ by $y \in Y$. Towards this end, we define a sub-relation L_{i} where $i \in \{M, O, A, D\}$ as an ordered triple (X, Y, G_{i}) for every different case to expand x. The letters for naming the sub-relation follow these different cases, i.e. expanding a cube by a *M*easure, *O*bject, *A*ttribute, or *D*imension.

Measure

This is a special case of the binary relation L that can be described as *"y is compatible to expand x by expanding the set of measures M_x"*. In this case, we define the binary relation L_{M} as an ordered triple (X, Y, G_{M}) where the sets of cubes $(x, y) \in G_{\mathrm{M}}$ with $G_{\mathrm{M}} \subseteq G$ fulfil the following requirements:

- $M_{\mathrm{y}} \backslash M_{\mathrm{x}} \neq \emptyset$ meaning $\exists z : z \in M_{\mathrm{y}}$ and $z \notin M_{\mathrm{x}}$
- $|D_{\mathrm{x}}| = |D_{\mathrm{y}}|$
- $\forall d_{\mathrm{i}} \in D_{\mathrm{x}} \exists! d_{\mathrm{j}} \in D_{\mathrm{y}} : (A_{\mathrm{i}} \subseteq A_{\mathrm{j}} | O_{\mathrm{i}} \subseteq O_{\mathrm{j}})$

At the simplest case, these requirements mean that cube y should have additional measures than x and the dimensions of y should have at least the objects of cube x.

Considering cube $x_1 \in X$ from the running example, cube $y_1 \in Y$ would be compatible to expand x by expanding the set of measures $M_{\mathrm{x_1}}$, in the case that $D_{\mathrm{y_1}} = \{geo_{\mathrm{y}}', time_{\mathrm{y}}', sex_{\mathrm{y}}'\}$ and $M_{\mathrm{y_1}} = \{poverty\}$, whereas the geospatial dimension has one more attribute, i.e. $A_{\mathrm{geo_y'}} = \{city, country\}$, but the same objects at the country level, i.e. $O_{\mathrm{geo_y'}} = \{BE, GR, IT, (Brussels, BE), (Athens, GR), (Rome, IT), (Milan, IT)\}$. We also consider that the other dimensions are the same, i.e. $time_{\mathrm{y}}' = time$ and $sex_{\mathrm{y}}' = sex$.

Object

This is a special case of the binary relation L that can be described as *"y is compatible to expand x by expanding the set of objects of an attribute $a_{i,j}$ of a dimension $d_i \in D_x$*. In this case, we define the binary relation L_{O} as an ordered triple (X, Y, G_{O}) where the sets of cubes $(x, y) \in G_{\mathrm{O}}$ with $G_{\mathrm{O}} \subseteq G$ fulfill the following requirements:

- $M_{\mathrm{x}} \subseteq M_{\mathrm{y}}$
- $|D_{\mathrm{x}}| = |D_{\mathrm{y}}|$

- $\exists! a_{\mathrm{i,j}}$ in $d_{\mathrm{i}} \in D_{\mathrm{x}} : a_{\mathrm{i,j}} \backslash a_{\mathrm{k,m}} \neq \emptyset$ where $a_{\mathrm{k,m}}$ in $d_{\mathrm{k}} \in D_{\mathrm{y}}$
- $\forall d_{\mathrm{k}} \neq d_{\mathrm{i}} \in D_{\mathrm{x}} \exists d_{\mathrm{l}} \in D_{\mathrm{y}} : (A_{\mathrm{i}} \subseteq A_{\mathrm{l}} | O_{\mathrm{i}} \subseteq O_{\mathrm{l}})$

This set of requirements mean that at least one attribute of a dimension of cube y should have additional objects compared to one of the attributes of a dimensions of cube x, while all the other dimensions should be the same. In addition cube y should at least have the set of measures of x.

Considering again the same example, cube $y_2 \in Y$ would be compatible to expand x_1 by expanding the set of objects of the attribute O_{geo}, in the case that $M_{\mathrm{y_2}} = \{unemployment\}$, $D_{\mathrm{y_2}} = \{geo_{\mathrm{y}}'', time_{\mathrm{y}}'', sex_{\mathrm{y}}''\}$and $O_{\mathrm{geo_y}''} = \{GR, IT, IE\}$, whereas $time_{\mathrm{y}}'' = time$ and $sex_{\mathrm{y}}'' = sex$

Attribute

This is a special case of the binary relation L that can be described as *"y is compatible to expand x by expanding the set of attributes A_i of a dimension $d_i \in D_x$"*. In this case we define the binary relation L_{A} as an ordered triple (X, Y, G_{A}) where the sets of cubes $(x, y) \in G_{\mathrm{A}}$ with $G_{\mathrm{A}} \subseteq G$ fulfill the following requirements:

- $M_{\mathrm{x}} \subseteq M_{\mathrm{y}}$
- $|D_{\mathrm{x}}| = |D_{\mathrm{y}}|$
- $\exists! d_{\mathrm{i}} \in D_{\mathrm{x}} : (A_{\mathrm{j}} \backslash A_{\mathrm{i}} \neq \emptyset)$ and $d_{\mathrm{j}} \in D_{\mathrm{y}}$. The relative complement of $A_{\mathrm{i,x}}$ to A_{j} means that $\exists a_{\mathrm{j,k}} : a_{\mathrm{j,k}} \in A_{\mathrm{j}}$ and $a_{\mathrm{j,k}} \notin A_{\mathrm{i}}$
- $a_{\mathrm{j,k}} \neq \emptyset$ meaning that this attribute contains one or more objects.
- $\forall d_{\mathrm{k}} \neq d_{\mathrm{i}} \in D_{\mathrm{x}} \exists d_{\mathrm{l}} \in D_{\mathrm{y}} : (A_{\mathrm{i}} \subseteq A_{\mathrm{l}} | O_{\mathrm{i}} \subseteq O_{\mathrm{l}})$

These requirements denote that at least one dimension of cube y should have additional attributes compared to one of the dimensions of cube x, while all the other dimensions should be the same. In addition cube y should at least have the set of measures of x.

Following the same example from section 4.3, cube $y_3 \in Y$ would be compatible to expand x_1 in the case that $A_{\mathrm{time_y}''} = \{quarter, month\}$ and $M_{\mathrm{y_3}} = \{unemployment\}$, whereas all other dimensions are the same.

We should also note that in this case a cube y can be created by materialising cube x i.e., compute aggregations of cube x across a hierarchy. In particular, if cube x is populated with data for an attribute of a dimension and there is also a hierarchy in which this attribute is at a low level, then aggregation for other attributes at a higher level can be also computed and thus a new cube can be created. This new cube can now play the role of cube y that expands x. In the case, however, that we need data for a lower than our attribute level then this process cannot be performed and thus cubes on the Web should be discovered.

Dimension

This is a special case of the binary relation L that can be described as *"y is compatible to expand x by expanding the set of dimensions D_x"*. In this case we define the binary relation L_{D} as an ordered triple (X, Y, G_{D}) where the sets of cubes $(x, y) \in G_{\mathrm{D}}$ has to fulfill the following requirements:

- $M_{\mathrm{x}} \subseteq M_{\mathrm{y}}$
- $|D_{\mathrm{x}}| < |D_{\mathrm{y}}|$
- $\forall d_{\mathrm{i,x}} \in D_{\mathrm{x}} \exists d_{\mathrm{j,y}} \in D_{\mathrm{y}} : (A_{\mathrm{i,x}} = A_{\mathrm{j,y}} | O_{\mathrm{i,x}} = O_{\mathrm{j,y}})$

At the simplest case, cube y has one extra dimension in relation to cube x and all the other dimensions the same. Moreover, cube y should at least have the set of measures of x.

Following the same example, cube $y_4 \in Y$ would be compatible to expand x_1 in the case that $D_{\mathrm{y_4}} = \{geo, time, sex, agegroup\}$ and $M_{\mathrm{x_4}} = \{(unemployment)\}$.

In this case, materialisation of cube x, i.e. computation of aggregations across a dimension, can create new cubes that can be expanded by cube x.

Other properties of L

These sub-relations L_{i} are not transitive, meaning that if y_1 is compatible to expand x in a particular way and y_2 is compatible to extend y_1 in the same way, then y_2 is not always compatible to expand x in this same way. For example, if we have two cubes y_1 and y_2 that have exactly the same dimensions as cube x_1 from the example in section

4.3, while $M_{y_1} = \{crime\}$ and $M_{y_2} = \{unemployment\}$ then y_1 is compatible to expand x_1, y_2 is compatible to expand y_1 but y_2 is not compatible to expand x_1.

We should also say that G_i are disjoint i.e. $\bigcap_{i \in \{M,O,A,D\}} G_i = \emptyset$.

4.3.1.2 Expanding Compatible Cubes

Let $x \in X$ and $y \in Y$ be two cubes so that y is compatible to expand x, i.e. $(x, y) \in G$ according to the analysis of subsection 4.3.1.1. We define the operator E that expands a cube x using a cube y and produces a cube $z \in Z$. We say that operator E *"expands x by y"* and we specify that $E : G \rightarrow Z$ for all $(x, y) \in G$ in order to denote that the operator maps from G to Z. We could also say that this operator is similar to a left outer join of two cubes.

Operator E is also related to the way that is used to expand x. Therefore, for each of the subsets $G_i \subseteq G$ where $i \in \{M, O, A, D\}$ we define an operator E_i that maps from G_i to Z, i.e. $E_i : G_i \rightarrow Z$.

Measure

The operator $E_M : G_M \rightarrow Z$ expands a cube x by adding elements into the set of measures M_x.

In this case, cube $z \in Z$ is defined by the following:

- $M_z = (M_x \cup M_y)$
- $D_z = D_x$

Following the same example of section 4.3 and sub-section 4.3.1.1, cube z_1 will be the outcome of expanding x_1 by y_1. The set of measures for z_1 will be $M_{z_1} = \{unemployment, poverty\}$, while the dimensions will be $D_{z_1} = \{geo, time, sex\}$.

Object

The operator $E_I : G_I \rightarrow Z$ expands a cube x by adding one or more elements into the set of objects of an attribute $a_{i,j}$ of a dimension $d_i \in D_x$.

In this case, cube $z \in Z$ is defined by the following:

- $M_{\mathrm{z}} = M_{\mathrm{x}}$
- $\exists a_{\mathrm{p,q}}$ in $d_{\mathrm{p}} \in D_{\mathrm{z}} : a_{\mathrm{p,q}} = a_{\mathrm{i,j}} \cup a_{\mathrm{k,m}}$ where $a_{\mathrm{i,j}}$ in $d_{\mathrm{i}} \in D_{\mathrm{x}}$ and $a_{\mathrm{k,m}}$ in $d_{\mathrm{k}} \in D_{\mathrm{y}}$
- $\forall d_{\mathrm{r}} \in D_{\mathrm{z}} | d_{\mathrm{r}} \neq d_{\mathrm{p}} : D_{\mathrm{z}} = D_{\mathrm{x}}$

Following the same example cube, z_2 will be the outcome of expanding x_1 by y_2 and will be characterised by $M_{\mathrm{z}_2} = \{unemployment\}$, $D_{\mathrm{z}_2} = \{geo_{\mathrm{z}}'', time_{\mathrm{z}}'', sex_{\mathrm{z}}''\}$ and $O_{\mathrm{geo_z"}} = \{BE, GR, IT, IE\}$, whereas $time_{\mathrm{z}}'' = time$ and $sex_{\mathrm{z}}'' = sex$.

Attribute

The operator $E_{\mathrm{A}} : G_{\mathrm{A}} \rightarrow Z$ expands a cube x by adding elements into the set of objects of an attribute A_{x}.

In this case, cube $z \in Z$ is defined by the following:

- $M_{\mathrm{z}} = M_{\mathrm{x}}$
- $\exists d_{\mathrm{p}} \in D_{\mathrm{z}} : A_{\mathrm{p}} = A_{\mathrm{i}} \cup A_{\mathrm{j}}$ where $d_{\mathrm{i}} \in D_{\mathrm{x}}$ and $d_{\mathrm{j}} \in D_{\mathrm{y}}$
- $\forall d_{\mathrm{r}} \in D_{\mathrm{z}} | d_{\mathrm{r}} \neq d_{\mathrm{p}} : D_{\mathrm{z}} = D_{\mathrm{x}}$

Following the same example cube, z_3 will be the outcome of expanding x_1 by y_3 and will be characterised by $M_{\mathrm{z}_3} = \{unemployment\}$ and $A_{\mathrm{time_z"'}} = \{year, quarter, month\}$ whereas all other dimensions will be the same.

Dimension

The operator $E_{\mathrm{D}} : G_{\mathrm{D}} \rightarrow Z$ expands a cube x by adding elements into the set of dimensions D_{x}.

In this case, cube $z \in Z$ is defined by the following:

- $M_{\mathrm{z}} = M_{\mathrm{x}}$
- $D_{\mathrm{z}} = (D_{\mathrm{x}} \cup D_{\mathrm{y}})$

Following the same example cube, z_2 will be the outcome of expanding x_1 by y_2 and will be characterised by $M_{\mathrm{z}_2} = \{unemployment\}$ and $D_{\mathrm{Z}} = \{geo, time, sex, agegroup\}$

It is important to see whether $E(E(x, y_1), y_2) = E(x, E(y_1, y_2))$. Probably y_1 and y_2 should be mutual compatible to expand each other.

4.3.2 Linked Data Cubes

The RDF Data Cube (QB) vocabulary Cyganiak and Reynolds (2014) is a *W3C* standard for modelling data cubes as graphs and thus adhering to the RDF model and Linked Data principles. Centric class in the vocabulary is *qb:DataSet* that defines a cube. A cube has a *qb:DataStructureDefinition* that defines the structure of the cube and multiple *qb:Observation* that describe each cell of the cube. The structure is specified by the abstract *qb:ComponentProperty* class, which has three sub-classes, namely *qb:DimensionProperty*, *qb:MeasureProperty*, and *qb:AttributeProperty*. The first one defines the dimensions of the cube, the second the measured variables, while the third structural metadata such as the unit of measurement.

Usually the values of the components are populated using predefined code lists that might formulate hierarchies such as a geographic or administrative division. These code lists can be specified by using either the Simple Knowledge Organisation System (SKOS) Miles and Bechhofer (2009) vocabulary or the QB vocabulary. SKOS is a W3C standard used for expressing the basic structure and content of concept schemes such as thesauri, taxonomies, and classification schemes. The set of values is modelled as a *skos:ConceptScheme* and a value as a *skos:Concept*. In addition, *skos:broader* and *skos:narrower* are used to assert a direct hierarchical link between two *skos:Concepts*. In case of reusing RDF data that are not modelled using SKOS, the QB vocabulary introduced the *qb:HierarchicalCodeList* class that defines a set of root concepts in the hierarchy (*qb:hierarchyRoot*) and a parent-to-child relationship (*qb:parentChildProperty*).

XKOS[1] RDF vocabulary has been proposed as an extension to SKOS that allows to model hierarchies structured in levels. A hierarchy level can be defined using the *xkos:ClassificationLevel* concept. According to XKOS the levels of a hierarchy are organised as an *rdf:List*, which implies order, starting with the most aggregated level. Individual *skos:Concept* objects are related to the *xkos:ClassificationLevel* to which they belong by the *skos:member* property.

[1] http://rdf-vocabulary.ddialliance.org/xkos

TABLE 4.1: Mapping of Multidimensional Model to Linked Data vocabularies

Multidimensional model	*Linked Data vocabularies*
Dimension	*qb:DimensionProperty*
Hierarchy	*qb:HierarchicalCodeList or an rdf:List that is linked to a skos:ConceptScheme through the xkos:levels property*
Dimension levels	*xkos:ClassificationLevel*
Dimension values	*Individual skos:Concept objects that are related to the xkos:ClassificationLevel to which they belong by the skos:member property.*
Measure	*qb:MeasureProperty*

The linking in linked data cubes is currently performed through the reuse of resources that define common statistical concepts and associated code lists that are used across multiple datasets. For example, the *sdmx-dimension:timePeriod* property is usually used for time, the *sdmx-dimension:refArea* for geography, and *sdmx-dimension:sex* for sex dimension.

The conceptual relations between the concepts that define a cube and the classes or properties of popular Linked Data vocabularies are summarised in Table 4.1.

However, when dealing with real world data cubes this conceptual mapping is not always straightforward. For example, a dimension attribute value has a conceptual relation to the *skos:Concept* of a *skos:ConceptualScheme*. However, if we need to find the attribute values of a dimension we have to query the actual *qb:Observations* and not the *skos:ConceptualScheme* as the latter could have more concepts that are not used in the specific cube.

During the last years a growing number of endeavours focused on linked data cubes.

Software tools have been developed to support both creation and exploitation of linked

data cubes. In the first case, the tools aim at transforming data from legacy technical formats ranging from CSV, JSON-stat and SDMX-ML to relational and OLAP databases into RDF data adhering to the RDF Data Cube (QB) vocabulary Capadisli et al. (2013); Kalampokis et al. (2014); Ruback et al. (2013); Salas et al. (2012a,b). In the case of exploitation, existing tools enable exploring cubes in two-dimensional tables and on maps, as well as creating charts Helmich et al. (2014); Kalampokis et al. (2014); Mader et al. (2014).

Moreover, a number of linked data cubes have been made available on the Web. Some of them have been created by official endeavours meaning that they have been launched by public bodies that own the data. For example, the European Commission's Digital Agenda provides its Scoreboard[2] as linked data cubes. Census data of 2011 from Ireland[3] and Italy[4] have been published as linked data by their National Statistics Institutes. The Department for Communities and Local Government (DCLG)[5] and the Scottish government[6] in the UK as well as the Flemish government[7] in Belgium opened up their statistics as linked data. For an in-depth description of these governmental endeavours see Appendix C. At the same time, a number of datasets have been transformed to linked data cubes in third parties activities. For example, a linked data transformation[8] of Eurostat's data, which was created in the course of a research project, includes more than 5,000 linked data cubes. Moreover, few statistical datasets from the European Central Bank, World Bank, UNESCO and other international organisations have been also transformed to linked data in a third party activity Capadisli et al. (2013). In addition, census data of 2011 from Greece Petrou et al. (2013) and historical censuses from the Netherlands Meroño-Peñuela et al. (2012) have been transformed to linked data.

During the last years, a few research endeavours focused on performing statistical analyses on top of combined linked data cubes Kalampokis et al. (2013c); Zapilko and Mathiak (2011); Zaveri et al. (2013). These endeavours mainly proposed ad-hoc solutions that use specific datasets in order to prove the applicability of the approach in specific domains. They are very important as they demonstrate the value of performing analytics

[2] http://digital-agenda-data.eu/data
[3] http://data.cso.ie
[4] http://datiopen.istat.it
[5] http://opendatacommunities.org/data
[6] http://statistics.gov.scot
[7] http://data.opendataforum.info
[8] http://eurostat.linked-statistics.org

on combined data sets on the Web but they do not propose a generic solution towards this end.

With regards to linked data cubes integration it was only recently when academia focused on the issue. Abello et al. Abello et al. (2014) explored how semantic web technologies can aid in data discovery, acquisition, integration, and analytical querying of external data, and thus serve as a foundation for OLAP exploration. Moreover, the *drill-across* operation Gómez et al. (2012) was recently extended in the context of linked data cubes Kämpgen et al. (2014).

Finally, various systems have been developed on top of linked data cubes, e.g. a question answering system Höffner and Lehmann (2014) and an access control framework and platform for medical data integration Kamateri et al. (2014).

4.3.2.1 A Process Model for Integrating Linked Data Cubes

According to the literature, open data processes specify the steps that governments should follow to set their data free for others to reuse. Moreover, a few processes have been recently proposed in the literature to describe the steps that are followed in publishing and consuming linked data. However, these processes are general and need to be specialised for accommodating statistical data modelled using linked data technologies. In particular, these generic processes present the following limitations when applied to linked data cubes:

- They focus on the publishing part of linked data and they do not provide details on the exploitation, which is usually summarised at the last step of the process. In our case, however, the possible statistical analyses are well defined in the literature (e.g. OLAP analysis, statistical learning etc.) and thus should be further elaborated particularly as they can also provide feedback to the publishing steps of the process.
- Typically, data integration in the Web of Linked Data is facilitated by establishing owl:sameAs links, which indicate that two URI references refer to the same thing. However, in the case of cubes these links are applicable only at the metadata level that define the structure of the cube and not at the observation level. As a result, integration of data cubes is not currently properly accommodated in existing linked data processes.

- The use of the QB vocabulary introduces considerable complexity that calls for specific requirements in the publishing steps.

This sub-section presents the proposed process for creating value through linked data cubes (Tambouris et al., 2015). In order to understand the requirements of a linked data cube process we interviewed employees from public and private organisations that work with open data, linked data, and statistical analysis. More specifically, the following appointments were made per area:

- Open data: The head of the open data team of the Flemish government.
- Linked data: Two employees from an international Swiss Bank.
- Statistical data: 16 employees of the Irish Central Statistics Office (CSO). The interviewees were chosen as a cross-section of CSO staff from different functional areas and different levels of seniority, with particular focus on staff involved in data dissemination and IT operations. One of CSO's major statistical datasets is Census 2011, which has been already published as Linked Data .
- Open/linked data: Three employees from the Research Centre of the Government of Flanders; a government having as mission statement to conduct research in the fields of demographics, macroeconomics and social-cultural developments.
- Open/linked data: Three employees from the Assistant Deputy Director of Strategic Statistics in the UK Department for Communities and Local Government (DCLG). DCLG currently produces 53 main statistical datasests and is commited to routiely release its data as linked open data. It also maintains a data portal that currently contains more than 150 datasets .

This process comprises three phases, namely (a) Creating Cubes, (b) Expanding Cubes, and (c) Exploiting Cubes. The first phase involves creating linked data cubes from raw data, the second supports the expansion of a cube by linking it with other cubes on the Web, and the last one enables the exploitation of the cubes in data analytics and visualisations. The three phases further split up into a number of steps. A depiction of this process is presented in Figure 4.12. In the rest of the section the steps of each phase are outlined.

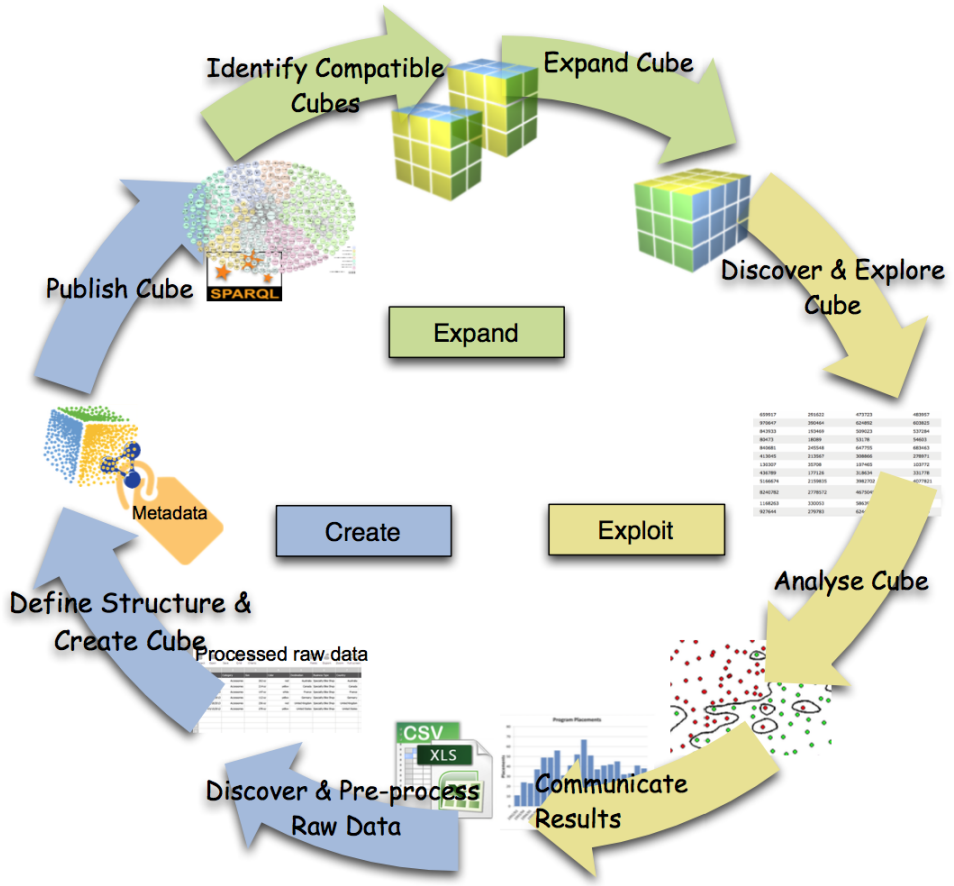


FIGURE 4.12: The linked open data cubes process (Tambouris et al., 2015)

Step 1.1: Discover & pre-process raw data

This step enables stakeholders to discover, access, view and process raw data cubes. At this step, data cubes come in various data formats such as CSV files, XLS files, RDBMS or RDF files. In addition, cubes can be formatted in various structures such as rectangular data, tree data and graph data.

In this step, stakeholders are able to browse raw data and perform activities aiming to improve the quality of raw data (e.g. data sorting, filtering, cleansing, transformation). This step could also include raw file or raw data storage in a local repository or database system. In this case, metadata regarding the provenance of raw data may be also stored along with the actual data.

Step 1.2: Define structure & create cube

An important step in linked data creation regards the definition of the structure of a model that the data will be mapped to. Initially, a conceptual model that drives the development of the structure of the linked data cube is created. This specifies:

- The dimensions of the cube, which define what the observation applies to.
- The measured variables (i.e. what has been measured) along with details on the unit of measure or how the observations are expressed.

As reusing widely accepted vocabularies is considered to be of high importance in linked data, defining the structure of the model also requires importing and reusing existing linked data vocabularies. In the case of data cubes, the RDF Data Cube (QB) vocabulary constitutes the main framework to model data cubes as RDF graphs. In addition, other linked data vocabularies can be also used to define the values of the dimensions, measures and attributes of the cube. Common statistical concepts can be reused across datasets e.g. dimensions regarding age, location, time, sex etc. or the values of specific dimension (e.g. the countries of Europe). These concepts are defined in linked data vocabularies that standardise dimensions, attributes and code lists. The most widely accepted is the SDMX-RDF vocabulary , which is based on the statistical encoding standard SDMX.

As a result, publishing linked data cubes mainly requires discovery and reuse of controlled vocabularies. We should also note that reusing controlled vocabularies could be considered as reconciling against such collections. This peculiarity of data cubes introduces an extra need that is related to the management of controlled vocabularies that could be reused across different datasets. This includes the creation, store, search, discovery and reuse of existing controlled vocabularies. This step also includes the creation of the actual RDF data out of the raw data based on the structure definition that was created at the previous step. This step includes the following activities: (a) URI design, (b) Definition of mapping between raw and RDF data, (c) Data storage to an RDF store, and (d) Validation for compliance with schema or values constraints.

Finally, this step also includes the enrichment of RDF data cubes with metadata to facilitate discovery and reuse. Sources of metadata include raw data files, the cube's structure and/or standard thesaurus of statistical concepts.

Step 1.3: Publish cube

In this step, the generated data cubes are made available to the public through different interfaces e.g. Linked Data API, SPARQL endpoint, downloadable dump etc. In addition, during this step the datasets are publicised in data catalogues such as Europe's public data portal or other national portals (e.g. data.gov.uk or data.gov.gr), the datahub platform or the Linking Open Data cloud.

Metadata that describe the dataset should be also published along with the actual data. The produced metadata are usually shared across multiple platforms and implementations. As a result, stakeholders need to be able to import or export metadata related to data cubes.

Step 2.1: Identify compatible cubes

This step supports the identification of compatible to join cubes in order to enable expanding linked data cubes. The identification of compatible cubes is performed through two processes:

- Search on an existing collection of linked data cubes and evaluate the compatibility of a cube at hand with every cube in the collection. The compatibility evaluation is based on (a) the structure of the cubes i.e. dimensions, measures, levels and hierarchies, and (b) the desired type of join. For example, a cube is compatible to join in order to add a new measure to an original cube if: i) both cubes have the same dimensions, ii) the second cube has at least the same values in each dimension of the original cube, and iii) the second cube has at least one measure that does not exist at the original cube.
- Create a set of compatible cubes from an initial linked data cube by computing aggregations across a dimension or a hierarchy. In the case of aggregating data across a dimension, 2n new cubes are created where n is the number of the dimensions of the cube. In the case of aggregating data across a hierarchy, a new cube is created that contains observations for all values of a dimension at every level. Special attention should be paid on the types of measures and dimensions and the aggregation function (i.e. sum, count, min, max etc.) that can be used.

This step can also include the establishment of typed links between compatible to join cubes. These links will enable, at a later stage, identifying linked data cubes that can be

combined in order to perform enhanced analytics on top of multiple linked data cubes. For this reason, it is important to define compatibility of cubes and develop tools that could search on large collections of cubes and discover cubes that can potentially be combined.

Step 2.2: Expand cube

Expanding cubes enables adding more data into a cube. We assume that a cube can be expanded by increasing the size of one of the sets that defines it. Therefore, a cube can be expanded by adding one or more elements into the set of measures, the set of concepts in a dimension and the set of dimensions. This can be done by merging a cube with a second one, which is compatible with the initial cube. The links that have been established at the previous step can be exploited towards this end.

Following the same example, we see that we have two cubes that describe two different measures (i.e. unemployed people and crime incidents) based on the same dimensions (i.e. time and geography), and with the same concepts (i.e. 2010 for time and the European countries for geography). These two cubes are compatible to merge and thus a new cube with two measures can be created out of the initial ones. We should note that the expanded cube could be either created and stored or just conceptually defined in order to be used along with a data analytics tool.

Step 3.1: Discover & explore cubes

At this step, stakeholders aiming to consume data exploit the mechanisms set up at the previous step in order to discover the appropriate cubes for a task at hand. For example, we consider a researcher that needs to study the relation between unemployment and criminality and thus needs to analyse data that describe unemployment and criminality in different geographic areas or time periods.

In general, the discovery of linked data cubes could be done through:

- A data catalogue that allows exploring the available data cubes based on (a) generic metadata records stored inside the catalogue platform that describe the cube as a whole, and (b) Cube-specific metadata that provide information about the concepts that formulate the cube i.e. dimensions and measures.

- Full-text search that enables discovery of data cubes not only by metadata but also by the actual content of the cubes. In our example, we suppose that the researcher identifies two cubes:
- A cube presenting the number of unemployed people in three dimensions i.e. countries, years and age groups.
- A cube presenting crime incidents in two dimensions i.e. countries and time (quarters of the year).

At this stage, we consider that the researcher is also able to browse the cube in order to better understand the data and proceed with further analysis. This enables the researcher to view data based on different dimensions or measures. For example, if the data describes the unemployment rate at different European countries in different years then stakeholders could view either the unemployment rates of a particular country throughout the years or the unemployment rates of a specific year across different countries. This would enable stakeholders also to sort or filter the data based on the values of the dimensions or the actual values of the observations.

Step 3.2: Analyse cube

In this step the data cubes that were resulted from the previous step are employed in order to perform analytics through (a) OLAP operations, (b) computing simple summaries of the data, and (c) creating statistical learning models. The transformation of linked data cubes at the previous step will enable stakeholders to perform the following OLAP operations:

- Dimension reduction: This would enable users to select part of a data cube by removing one of the dimensions. In the unemployment rate example this would enable, for example, removing the age group dimension and thus keeping only the time and location dimensions.
- Roll-up and drill-down operation: These OLAP operations allow stakeholders to navigate among levels of data cube by stepping down or up a concept hierarchy.

Following the previous example, stepping down a concept hierarchy for the dimension time could perform this OLAP operation. If we consider the concept hierarchy

?month¡quarter¡year? then drill down would present unemployment rate of different age groups at different countries for every quarter.

The stakeholder could also select to produce either quantitative (i.e. summary statistics) or visual (i.e. simple to understand graphs) summaries. As regards the quantitative summaries, a stakeholder in this step will be able to describe the observations across a dimension using descriptive statistics. For example, this step would enable the calculation of the mean and standard deviation of the unemployment rate of European countries in a particular year. Moreover, stakeholders would be able to calculate statistics (e.g. Pearson's correlation coefficient) that estimate dependences between paired measures described in disparate but compatible cubes. Paired here is used to denote that the measures share at least one common dimension and thus can be compared.

Finally, the types of visualisation charts that can be used in this step include scatter plots, bar charts, pie charts, histograms, geo charts, timelines etc.

Following the example of the previous steps, the researcher use the cubes created after the last step in order to perform the following:

- Create a scatter-plot presenting unemployed people against crime incidents across European countries.
- Calculate Pearson's correlation coefficient between number of unemployed and number of crimes.

In this step, the cubes that were created in the previous steps could be also used in machine learning and predictive analytics in order to produce learning or predictive models. At the same step, the models that were created could also be published into the Linked Data Web and thus feedback the lifecycle at the first step. Following the example of unemployment and criminality, we consider that the researcher now wants to create a model in order to be able to estimate future crime rates based on unemployment rates. Towards this end, the researcher exploits the results of the previous step and the data cubes in order to select an appropriate data mining method (e.g. Support Vector Machines) and build a model. The researcher goes back to the previous steps in order to also identify data to evaluate the model.

Step 3.3: Communicate results This step involves the visualisation of results. This step may feed back to the first step of the process if the results of the analyses performed in the previous steps indicate a need for further analyses requiring additional data. Towards this end, the analysis proceeds with the first step of the process in order to discover new raw data, transform them to RDF and eventually perform a comparative analysis with existing RDF data cubes.

4.3.2.2 The Case of Linked Data Eurostat

In this subsection we study an unofficial transformation of Eurostat's data in order to quantitatively evaluate the applicability of the proposed operators in real world settings. In specific, this stage aims at identifying the number of compatible cubes in Eurostat for two of the four proposed operators. The linked data transformation of Eurostat's data includes more than 5,000 cubes ranging from 94 KB to 22 GB in size and with a median of 3 MB and average of 118 MB. The data structure definitions along with the code lists of the cubes are provided through a SPARQL endpoint while the actual data through one RDF file per cube. Moreover, although the cubes are modelled based on the QB vocabulary, the following practices can be find in the dataset:

- Measures are defined using *sdmx-measure:obsValue* that is declared as a *qb:DimensionProperty*.
- In cubes with multiple measures an extra *qb:DimensionProperty* is defined.
- Attributes such as frequency and unit are defined as *qb:DimensionProperty*.

Finally, the code lists, which are defined as *skos:ConceptScheme*s, are flat meaning that the *skos:Concept*s are not grouped into levels.

The fact that the observations are not provided through an endpoint hinders identification of the number of compatible cubes through a small number of SPARQL queries. To overcome this situation we introduce the concept of cluster and we analyze a set of clusters instead of the whole dataset. We define a cluster as a set of cubes C_i with i=$\{1...n\}$, where n is the size of the cluster, and $D_j = D_k \; \forall$ j and k $\in$ i. Hence, a cluster comprises cubes having the same dimensions. Moreover, we define the overlap of a dimension D_m that is common in all cubes in a cluster as the average AVG($|I_{km} \cap I_{jm}|/|I_{km} \cup I_{jm}|$) $\forall$ j,k $\in$ i and j $\neq$ k. In this context, we perform the following steps:

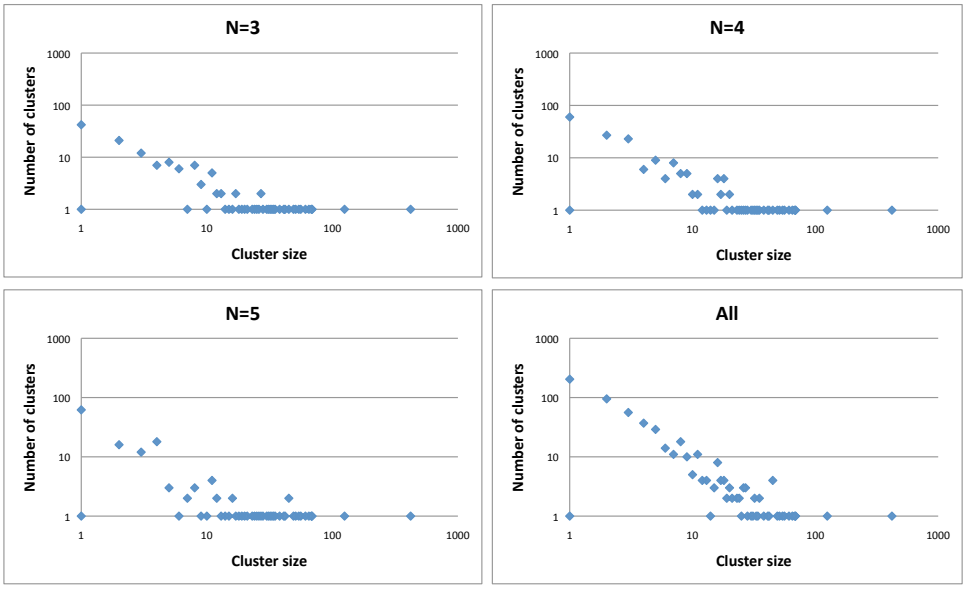


FIGURE 4.13: Log distribution of clusters of cubes having the same dimensions and the number of cubes formulating the cluster for cubes with N dimensions.

1. Identify clusters of cubes in the linked data version of Eurostat. We use the data structure definitions through the SPARQL endpoint to identify cubes of which the URIs of their dimensions or of the code list they use are identical.

2. Analyse the identified clusters. We select 10 clusters, import their dump files into an RDF store, and calculate the overlap of each dimension type along with the overlap of the measures in the cluster.

3. Identify the number of compatible cubes per cluster for every proposed operator.

Identify Clusters

In total we identified 562 clusters of cubes in the Linked Data version of Eurostat, while their size (i.e. the number of cubes inside a cluster) ranges from 2 to 421 cubes. The number of the dimensions of the cubes formulating a cluster ranges from 2 to 7 dimensions. Figure 4.13 presents the distribution of these clusters and the size of the clusters for cubes with different number of dimensions. The X-axis shows the size of a cluster in the log scale, while the Y-axis represents the corresponding frequency (in the log scale) of clusters of a specific size. Figure 4.13 comprises three diagrams for cubes with N=3, N=4, and N=5 dimensions and one diagram for all the cubes. We can observe that these four diagrams follow the same pattern, which is close to a Zipfian distribution, with a few clusters comprising a large number of cubes with the same dimensions.

TABLE 4.2: Analysis of clusters in the Linked Data version of Eurostat

No.	*Size*	*D1 overlap*	*D2 overlap*	*D3 overlap*	*D4 overlap*	*D5 overlap*	*M overlap*
1	69	46%	100%	-	-	-	2%
2	45	71%	59%	-	-	-	0.5%
3	34	45%	40%	8%	-	-	5%
4	15	60%	74%	15%	-	-	32%
5	37	84%	59%	43%	-	-	1.5%
6	20	90%	70%	22%	68%	-	19%
7	30	15%	85%	59%	18%	-	59%
8	21	99%	94%	99%	100%	-	0%
9	31	70%	31%	62%	89%	48%	2%
10	16	94%	86%	100%	91%	0.3%	100%

Analyze Clusters

Table 4.2 describes 10 clusters based on (a) their size, (b) the dimensions of the cubes in the cluster, (c) the dimension overlap per dimension type, and (d) the measure overlap. To ensure clarity we assume that always dimension D1 is the one that is related to geography, while D2 the one related to time.

Identify Compatible Cubes

Table 4.3 describes the number of compatible cubes per proposed operator that were identified in each of the 10 clusters.

4.3.2.3 Challenges in Integrating Linked Data Cubes

However, the implementation of the theoretical framework in the linked data realm encounters a number of challenges. In order to identify these challenges we study OGD endeavours that open up aggregated multi-dimensional data by exploiting linked data paradigm. In particular, we study six OGD portals, namely the Scottish government, the UK Department for Communities and Local Government (DCLG), the Flemish

TABLE 4.3: Number of compatible cubes per operator that were identified in each of the 10 clusters

No.	*Measure*	*Object*
1	1780	45
2	433	38
3	35	13
4	20	31
5	278	13
6	14	27
7	13	150
8	364	0
9	76	0
10	0	146

government, the Irish CSO, the Italian national statistics office, and Digital Agenda. These portal are described in detail in Appendix C.

We have categorised the challenges that we identified as follows:

- Challenges related to the different practices that can be followed in applying the QB vocabulary.
- Challenges related to the misuse of the QB vocabulary.
- Challenges related to the re-use of controlled vocabularies and code lists.
- Challenges related to use of proposed extensions of the QB vocabulary.
- Challenges related to conceptual issues.

Different practices in applying QB vocabulary

In many cases, the flexibility of the QB vocabulary enables publishers to follow different practices for publishing linked data cubes. These different practices hampers (a) the development of generic tools that can be used across different linked data cubes as well as (b) the combination of cubes across multiple sources.

Here the most important challenges are related to the understanding of the semantics of a measure. A widely adopted practice when referring to a qb:MeasureProperty is to use sdmx-measure:obsValue. For example Digital Agenda uses sdmx-measure:obsValue but it also defines an indicator dimension for which values are measured. The indicator takes values from a code list. In addition, DCLG, the Flemish government and the Irish CSO define measures as rdfs:subPropertyOf sdmx-measure:obsValue. DCLG data generally uses measure properties with quite specific semantics. In other data collections, Swirrl has used a small number of more generic measure properties, such as 'count' and 'ratio', defined as subproperties of sdmx-measure:obsValue. These work in conjunction with a sdmx-attribute:unitMeasure which defines what the observation is a 'count' of for example people or households. These unitMeasure values are re-used across datasets wherever possible to maximise the opportunities for combining and comparing data.

Often multiple measures need to be included in a data cube. The QB vocabulary proposes two approaches to include multiple measures in data cubes: (a) multimeasure observation or (b) qb:measureType. In the first approach the multiple measures can be declared as qb:MeasureProperty components in the structure of the cube. Each observation can be then attached with multiple observed values. One problem with this approach is that it allows the attachment of only a single attribute to each observation that will describe only one of the measured values. This could be fixed using the qb:componentAttachment property so as to attach one attribute to each qb:MeasureProperty but this attachment will regard the whole data set and can not vary between observations. The qb:measureType approach overcomes the previous problems. More precisely the second approach suggests to add extra dimensions to the structure of the cube using the qb:measureType component. These extra dimensions will actually play the role of the measures of the cube. Each observation of the cube will then have a single measured value. The disadvantage of this approach is that it substantially multiplies the number of triples potentially leading to performance and storing issues in the triple store that are stored. It is difficult to create a generic tool that consumes data following both approaches. The Irish CSO and Digital Agenda currently do not use multiple measures while DCLG uses the qb:measureType property option. Finally, the Flemish government employs both approaches. However the qb:measureType approach seems to be the most extensible and flexible one, due to the fact that it allows the use of much metadata/attributes for every individual observation as needed.

The QB vocabulary offers the possibility to group a set of observations into a slice where all but one or a small number of dimensions are fixed. The slice offers a mechanism for attaching metadata to that group of observations. The main example data collections examined in this chapter, from DCLG, Irish CSO, Flemish Government and EU Digital Agenda, do not currently make use of slices. However, recent developments in our approach (mainly related to the browser developed in PublishMyData environment) have found slices beneficial in two main respects. Firstly, for large data cubes, selecting observations to display in a two-dimensional table can lead to SPARQL queries that are expensive to execute. If observations are already associated with two-dimensional slices, this provides a convenient index that simplifies and speeds up such queries. Secondly, for data cubes with many dimensions, it is often the case in practice that these cubes can be sparse some combinations of dimension values do not have associated observations. In this case, it can sometimes be difficult for a user to navigate to populated parts of the cube. A user interface can present the user with a list of slices as a way of simplifying navigation to interesting, popular, or simply non-empty combinations of dimensions.

Moreover, the QB vocabulary allows two different practices for defining the allowed values of a dimension within a data set: (a) by connecting the qb:ComponentProperty with a qb:codeList property or (b) by defining the qb:ComponentProperty with a range of skos:Concept or a subclass of skos:Concept. Digital Agenda for example follows the first approach and connects qb:DimensionProperty with a codelist (it uses codelist `http://eurostat.linked-statistics.org/dic/geo#` from Eurostat). DCLG and Irish CSO in most cases do not associate dimension properties with a specific codelist but defines them with a range of skos:Concept, or a more specific class which is a subclass of skos:Concept. For example, the Irish CSO preserves data about 12 geographical hierarchical levels and defines a different concept scheme per geo level. This practice may be convenient but impedes the computation of aggregations as a complete codelist with levels is required.

QB vocabulary misuse

There are a few cases where the creation of linked data cubes is not consistent with what the QB vocabulary specifies. In such cases it is difficult to reuse generic tools for either exploiting or expanding data cubes.

For example, the RDF Data Cube vocabulary suggests the use of one qb:DimensionProperty for each of the cubes dimensions. Digital Agenda follows a very particular approach for the definition of its data sets' dimensions where a "super-dimension" is defined to embrace the values of dimensions other than time and location. Precisely, a "super-dimension" named "breakdown" is used to represent several values of dimensions including, for example, dimensions labeled as "Individuals who are born in non-EU country", "Individuals with high formal education" or "Unemployed". This approach facilitates the creation of RDF out of a huge data warehouse with hundreds of dimensions. However this "super-dimension" approach also generates problems in (a) developing generic tools that consume RDF data cubes, and (b) combining data cubes.

Re-use of controlled vocabularies and code lists

It is very important in linked data cubes to follow the main principle of linked data and re-use whenever possible existing URIs that describe resources or classes and properties. This should be happened to defined dimensions, objects of dimensions, levels of dimensions, measures, unit of measures, etc. This is of great importance for the combination of different data cubes. If different but related concept schemes are used, it is important to be able to define relationships between them.

For example, the time dimension is very important in most data cubes. A common approach for the time dimension property of a cube is to use sdmx- dimension:timePeriod or sdmx-dimension:refPeriod or a subproperty of them. For example, Digital Agenda uses the `http://semantic.digital-agenda-data.eu/def/property/time-period` property, a subproperty of sdmx-dimension:timePeriod. Moreover, DCLG uses a subProperty of `http://purl.org/linked-data/sdmx/2009/dimension#refPeriod` that is defined to have a range of `http://reference.data.gov.uk/def/intervals/Interval`. Finally, the Irish CSO does not use a time dimension in most of its data sets. However, when it does, it employs a resource of its own the name of which derives from the specific dataset (e.g. `http://data.cso.ie/census-2011/property/household-year-built`). Regarding the values of the time dimension of a cube, two different approaches are also used: (a) employing a predefined URI or (b) employing a literal value. For example the year 2014 could described as a resource e.g. `http://reference.data.gov.uk/id/gregorian-year/2014` or as a literal '2014'. DCLG and Digital Agenda standardises on URIs for time intervals provided by reference.data.gov.uk. These are clearly defined

with start and end points to the time interval and allows use of commonly occurring but reasonably complex intervals such as 'government years' which in UK run from 1 April to 31 March. The use of URIs offers more precise definitions of the time interval as long as these URIs are predefined and provides with all the linked data advantages such as the facilitation of the identification, linking or comparability with other cubes. A challenge here is the ability to correctly order the values of the time dimension in time and not in lexical order. Regarding the second approach, the use of literal values for the time dimension facilitates the SPARQL querying of the cube using, for example, queries such as "select observations made before 2014" or "select the most recent observation".

A geospatial dimension is also of high importance in most data cubes. A standard approach to define the geospatial dimension of a cube is to use sdmx-dimension:refArea property or a sub-property of the sdmx-dimension:refArea property. The Irish CSO for example uses sdmx-dimension:refArea property for the geospatial dimension. On the contrary, Digital Agenda uses http://semantic.digital-agenda-data.eu/def/property/ref-area which is sub-property of sdmx-dimension:refArea and DCLG uses http://opendatacommunities.org/def/ontology/geography/refArea also a sub-property of sdmx-dimension:refArea.

There is currently a need for constructing a commonly accepted code list for the units of measures of cubes. The code list will embrace the different units of measurements and be reused by different data sets. The lack of such commonly accepted code list results in the adoption of different code lists for the unit values in different data sets. For example Digital Agenda uses units from a code list of its own (http://semantic.digital-agenda-data.eu/codelist/unit-measure). In addition, DCLG and the flemish government use QUDT (http://www.linkedmodel.org/doc/qudt-vocab-units/1.1/index.html) which facilitates the conversion to other units. DCLG also uses DBpedia for currencies, and in particular http://dbpedia.org/resource/Pound_sterling. Finally, the Irish CSO does not define units for measures at all.

Finally, the definition of machine-readable hierarchical relationships (existing e.g. in geospatial data) is very useful for enabling aggregations within code lists. Nevertheless such relationships are generally not widespread in code lists. For example the Irish Census and Digital Agenda do not define hierarchies. DCLG also does not currently define hierarchies within code lists although its data sets include geographical

hierarchies. There are currently two approaches for defining hierarchical relationships: (a) using qb:HierarchicalCodeList or (b) adopting the SKOS or XKOS vocabularies. The qb:HierarchicalCodeList is introduced by the RDF Data Cube vocabulary and defines a set of root concepts in the hierarchy (qb:hierarchyRoot) and a parent-to-child relationship (qb:parentChildProperty). The SKOS vocabulary offers skos:broader and skos:narrower properties to enable the representation of hierarchical links. Moreover, XKOS, an extension of SKOS, also allows the modelling of hierarchies structured in levels. A hierarchy level can be defined using the xkos:ClassificationLevel concept. According to XKOS the levels of a hierarchy are organised as an rdf:List, which implies order, starting with the most aggregated level. Individual skos:Concept objects are related to the xkos:ClassificationLevel to which they belong by the skos:member property. Although XKOS seems to be a promising solution for the definition of machine-readable relationships, it is not currently commonly used.

Conceptual issues

An important challenge that hampers the development of tools that combine data cubes across the Web is the granularity of the cube. Different publishers specify cubes of different size. For example, the Irish Census of 2011 has defined 682 linked data cubes with one measure per cube while Digital Agenda only 2 cubes with more than 100 measures per cube. In such cases different approaches need to be followed in order to integrate data from two cubes and exploit them.

Chapter 5

Exploit Open Data in Analytics

5.1 Introduction

This chapter explores how integrated open data can be exploited in *data analytics* in order to create added value. This chapter focuses on open data structured in a multi-dimensional manner and considers two types of analysis:

- *OLAP analysis* (section 5.3). In this case, the potential of performing OLAP operations on top of integrated views of multiple data sets on the Linked Data Web is explored. Towards this end, we describe an innovative linked data OLAP browser that we have developed and we demonstrate its functionalities by combining and analysing OGD published by two governments in Europe, namely the Flemish and the Scottish governments.

- *Predictive analysis* (section 5.4): In this case, we focus on predictive analytics as a promising way of exploiting data on the Web. We give particular emphasis on Social Media data because they incorporates personal opinions, thoughts, and behaviours making them a vital component of the Web and fertile ground for a variety of business and research endeavours. In particular, we initially explore the predictive power of SM and define a process model for performing predictive analytics by exploiting data on the Web. Based on these, we design a case study that aims at predicting the winning party of UK elections 2010 by exploiting combined Twitter and linked open data.

5.2 The concept of Linked Open Government Data Analytics

The difficulty in exploiting open data seems surprising if we consider the huge importance data have in modern societies. Indeed, during the last years, businesses, academia and government employ various data analytics methods on their own data with great success. For example, business intelligence methods are employed by enterprises to help them survive in the global economy. In addition, evidence based policy-making relies on data analytics to assist policy makers in producing better policies. Finally, academia employ data analytics to test hypotheses, understand patterns, predict future points, estimate hidden parameters etc. in various domains and problem areas. We claim that the real value of OGD will unveil from performing data analytics on top of combined statistical datasets that were previously closed in disparate sources and can now be linked in order to provide unexpected and unexplored insights into different domains and problem areas. For this purpose, we deem that the linked data paradigm must be first adopted for constructing the technical infrastructure that is essential for employing data analytics in a decentralised manner on the Web.

A big portion of Open Government Data (OGD) concerns statistics such as population figures, economic and social indicators. Major providers of statistics on the international level include Eurostat, World Bank, OECD and CIA's World Factbook. Moreover, public agencies at all administrative levels collect, produce and disseminate statistical data through their OGD portals. Accurate and reliable statistics provide the solid ground for developing models that could support academia to better understand the world and businesses to make better decisions. These models enable the identification of patterns, prediction of future points and estimation of hidden parameters.

The availability of accurate and reliable statistical OGD in formats that enable easy reuse and combination can provide new potentials to businesses, academia and governments. The combination of statistical OGD that refer to different domains and is published by different public authorities with other data (e.g. enterprise's own data) could enable creating and evaluating models that were previously hard or even impossible to develop.

The potential of performing analytics on top of combined OGD and third party data could be summarised in the following user stories:

"As a business manager I want to be able to combine enterprise's data with accurate and timeliness demographics, economic and social indicators in order to make better decision regarding business operations and strategies". For instance, the correlation of product sales with economic and social indicators in various locations can reveal valuable information regarding consumer behaviour, hence supporting marketing or logistic departments. "As a researcher I want to be able to combine statistical data from disparate sources and domains in order to empirically identify novel hypotheses or test existing ones with more data as well as to understand patterns, predict values and estimate hidden parameters". For instance, developing models that integrate biodiversity information from a variety of datasets to assess biodiversity change, including remote sensing and in situ observations. "As a policy maker I want to be able to combine statistical data regarding economic and social indicators in order to identify evidences regarding policy interventions and hence evaluate policies." For instance, the correlation of data about education, unemployment and criminality in different geographical or administrative units and different time intervals could support or challenge existing policies.

However, putting together statistics in a meaningful manner so that to enable the creation of added value is usually a labour intensive task that introduces significant burdens to data users. It requires the manual discovery, collection, cleaning, transformation, integration, visualisation and statistical analysis of data. The vision that we present in this thesis suggests shifting this effort from the end-users to the data providers enabling this way the easier and wider reuse of OGD in various problem areas. As a result, OGD will be openly available for reuse in a way that will facilitate the performance of data analytics on top of combined open data and thus will enable the creation of useful information in an easy and cost effective manner.

For this purpose, the data should be provided in such a way that facilitates the whole lifecycle of statistical data reuse:

Data discovery: Metadata that describes statistical data should facilitate the effective and easy identification of datasets that could be combined for statistical analysis. This includes the identification of datasets that share common joint points (i.e. parliament constituencies, local authorities, schools etc.) and thus allow for further analysis. For instance, it is not feasible to correlate schools' expenditures with hospitals' inpatients

because there are no joint points between them. In addition, it includes the identification of datasets that describe variables measured using similar categories of units e.g. continuous or discrete. Finally, the metadata should enable the identification of variables of a specific category or class.

Data cleaning: The statistical data should be of high quality i.e. timely, accurate and relevant.

Data linking: The data should be linked in order to enable analysis in different levels of granularity e.g. unemployment that refer to parliamentary constituencies' level with criminality that refer to local authorities' level. Data linking should also facilitate the disambiguation of entities, concepts, units, codes etc. that are described in the datasets.

Data visualisation and statistical analysis: The data should enable easy visualisation and statistical analysis. Towards this end, the provided data should facilitate the automatic identification and matching of the unit of measurement of the described variables. This will allow the automatic visualisation and selection of the method to be used for the statistical analysis. For example, in the case of continuous units data analytics could be performed through linear regression analysis while in the case of discrete unit (i.e. categorical measures) through a classification analysis method such as logistic regression.

We should note, however, that the idea of Linked Open Government Data (LOGD) Analytics is different from traditional statistical analysis. LOGD Analytics are exploratory and thus they are not guided by theory. They do not provide any interpretation of the results and simply detect pattern and correlations. This is similar to what is often referred to as "the end of theory" due to data deluge (Anderson, 2008). The reader is kindly invited to speculate about potential theories behind these correlations, being aware that such speculations can often be wrong. For example, machines learned from Big Data that orange used cars have the best-kept engines, that passengers who preorder vegetarian meals usually make their flights, and that spikes in the sale of prepaid phone cards can predict the location of impending massacres in the Congo (Hardy, 2012).

5.2.1 Exploring the UK Elections through Open Data

In this sub-section we employ a simple use case in order to demonstrate the end-user value of the proposed LOGD analytics approach (Kalampokis et al., 2013c). In particular, our aim is to demonstrate how one can gain insights about UK elections from available Open Government Data (OGD) that is published on data.gov.uk. Towards this end, we use as a starting point open data regarding the outcome of two UK general elections from 2005 and 2010. In particular we employ datasets that are published on Guardian's web site under an open license regarding the elections results. In particular, these datasets contain the final results of all participating political parties in the country but also results of the main parties per parliamentary constituency along with the winning party per constituency. These datasets from Guardian's web sites were published as spreadsheets.

In order to be able to perform data analytics on top of combined datasets we need to identify datasets that share the same dimensions. In the elections case, we need to identify datasets with statistics on data.gov.uk that describe data in the parliament constituency level. Towards this end, we searched and collected datasets regarding unemployment. We have identified two datasets describing the unemployment and employment rate while we identified datasets in different time periods from 2005 until 2010. These datasets was published on data.gov.uk as spreadsheets.

Before we proceed with the actual analysis, the data goes through a data conditioning phase in which all the datasets from both data.gov.uk and Guardian are being published following the linked data principles and based on the technical requirements that we have specified before.

We have created an RDF data cube with unemployment rate as measure and year and parliament constituencies as dimensions. We have also published the elections datasets as linked data and according to the same requirements. In this case we have as a measure the number of votes and as dimensions the years, parliament constituencies and political parties. We have also created a cube that describes the winning party as measure and parliament constituencies and year as dimensions. In order to disambiguate the dimensions we have created typed links between our datasets and DBpedia, the linked data version of Wikipedia. We have used owl:sameAs links in order to denote that an

entity described in our dataset is the same as the corresponding entity in DBpedia. In this way, it becomes possible to disambiguate entity instances.

Two types of links are being established. The first one refers to the dimensions level while the second one to the observation level. Figure 5.1 presents the structure of a small part of the RDF data representing the data cubes along with the links among them. In particular, it describes the data structure definition of the unemployment rate and the election results cubes along with one observation per cube. It also presents the links that have been established between the two cubes. At the dimension level the graph denotes that one of the dimensions in both cubes is same as the dbpedia:United_Kingdom_constituencies entity while at the observation level that one observation in both cubes refers to an entity that is same as the dbpedia:Kensington_(UK_Parliament_constituency) entity. The final task is to annotate the described measure with a set of categories. In our case the elections datasets could be categorized under elections while the data.gov.uk datasets are being published based a predefined set of categories from the Office of National Statistics.

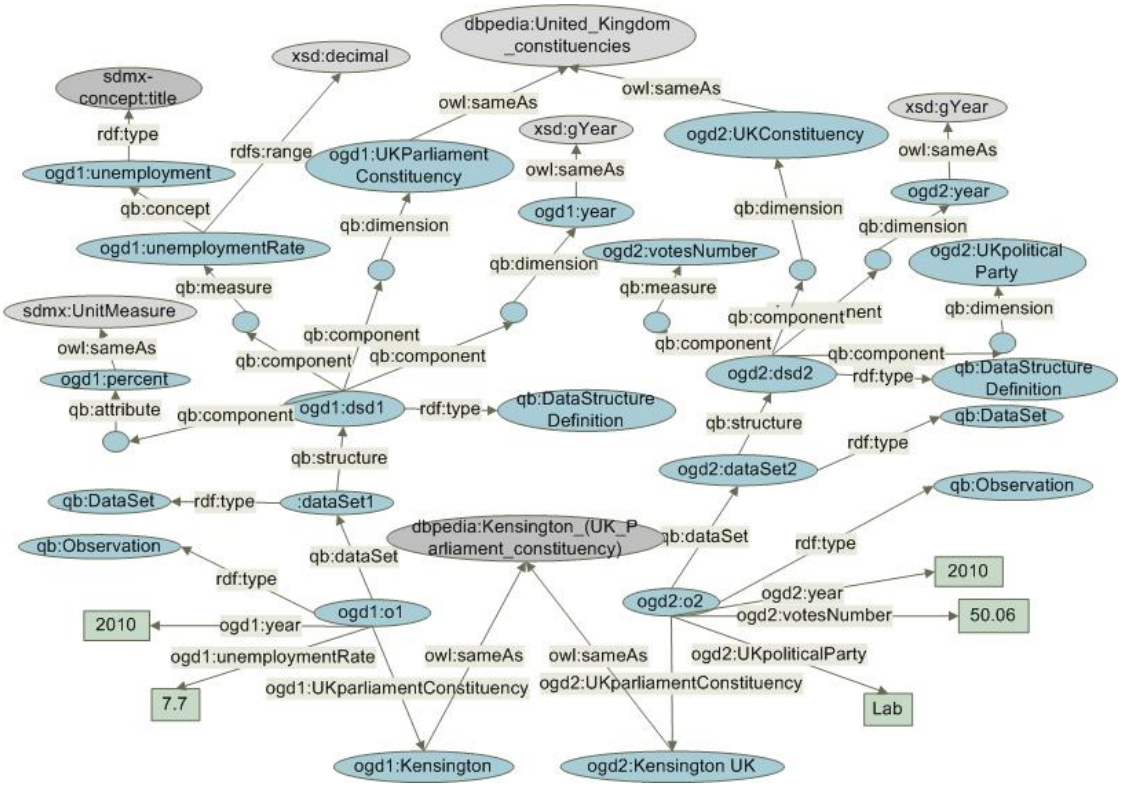


FIGURE 5.1: A part of the RDF graph that presents two data structure definitions of two cubes along with two observations and the respective links between them.

The data conditioning process gets the data into a state where data analytics can be performed in a transparent and easy manner. Initially, data that is related to elections can be discovered based on the category annotation. In our case scenario, two datasets that describe the number of votes and the winning party are discovered. Thereafter, datasets that share at least one dimension with the elections dataset are identified through the typed links that have been established to other datasets. In our case these datasets describe unemployment.

In addition, the datasets describe that the measures has been computed in either continuous (i.e. number of votes) or discrete (i.e. winning party) units. So, the way of visualizing the data or even the statistical analysis method to be followed can be emerged from the unit that characterizes the measure. For example, in the case of continuous units data analytics could be performed through linear regression analysis while in the case of discrete unit (i.e. categorical measures) through a classification analysis method such as logistic regression. Logistic regression measures the relationship between a categorical dependent variable and one or more continuous independent variables, by converting the dependent variable to probability scores through the logistic function:

$$P(A) = \frac{1}{1 + e^{-y}}$$

For example, in the elections case the logistic function could correlate the unemployment rate of a parliament constituency to the probability P(A) a particular political party to win the elections in the same constituency. On the other hand, the linear regression would correlate the unemployment rate of a parliament constituency to the number of votes that a specific political party received in the same constituency.

In our case we assume that the elections related dataset that includes a continuous measure is analysed based on a linear regression while the one that includes a discrete measure based on logistic regression. The analysis can be also combined with a comparative visualisation depending on the selected analysis method. The comparative visualisation enables the easy understanding of the data analytics results and facilitates their interpretation.

In Figures 5.2 and 5.3 the visualisations of linked open government data analytics are depicted. These figures visualise logistic regression analyses for the winners per constituency datasets in relation to the unemployment rate per constituency datasets for two consecutive UK elections in 2005 and 2010. In particular, the visualisations depict the percent of constituencies of a particular unemployment rate in which a party has won the elections. For example, in Figure 2 in the diagram that refers to the 2005 elections we see that the conservatives did not win in any parliament constituency with ten percent (10%) unemployment rate. In addition, as the same figure indicates for parliament constituencies with more than five percent (5%) unemployment rate the Conservative Party has very small probability to win. However, in the same Figure and in the diagram that refers to the 2010 elections the unemployment rate above of which the Conservative party has very small probability to win goes up to thirteen percent (13%).

These visualisations enable users to evaluate the performed analyses and thus to understand the correlation between measures described in different datasets. As Figures 5.2 and 5.3 suggest there is a significant relationship between the probability one of the two main UK parties (i.e. Labour party and Conservatives) to win in a parliament constituency and the unemployment rate in the same parliament constituency. It is notable that the same patterns holds for the two consecutive elections. The Conservative Party seems to win in areas that are characterised by small unemployment rate while the Labour Party in areas with high unemployment rate. Here we should note that the average unemployment rate in 2005 was 3.35 percent while in 2010 7.5 percent. This difference in the average unemployment rate between 2005 and 2010 could explain the moving of the data points and the data pattern to the right in the case of the unemployment measure. Finally, we should note that other connections that could have been produced by visualising or correlating other measures (e.g. poverty) are not presented in this chapter for shortness.

The identified patterns as presented through visualisation could be valuable for supporting decision-making. For example, political parties and candidates could intensify their campaigns in areas that the analytics predict negative results.

The presented approach also enables the creation of a statistical model out of the identified datasets. Towards this end, we used R statistical package and we computed the coefficients of a logistic function for four consecutive general elections from 2001 until

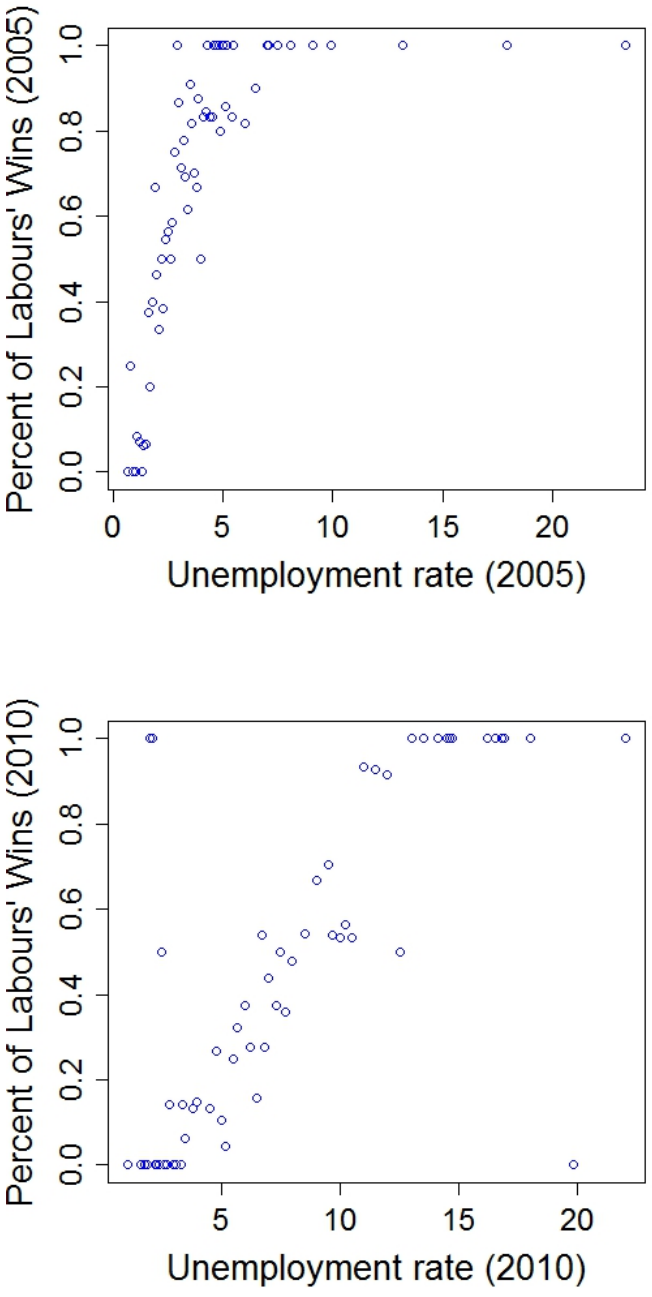


FIGURE 5.2: The correlation of Labours' wins and unemployment rate for the general elections of 2005 and 2010 (Kalampokis et al., 2013c)

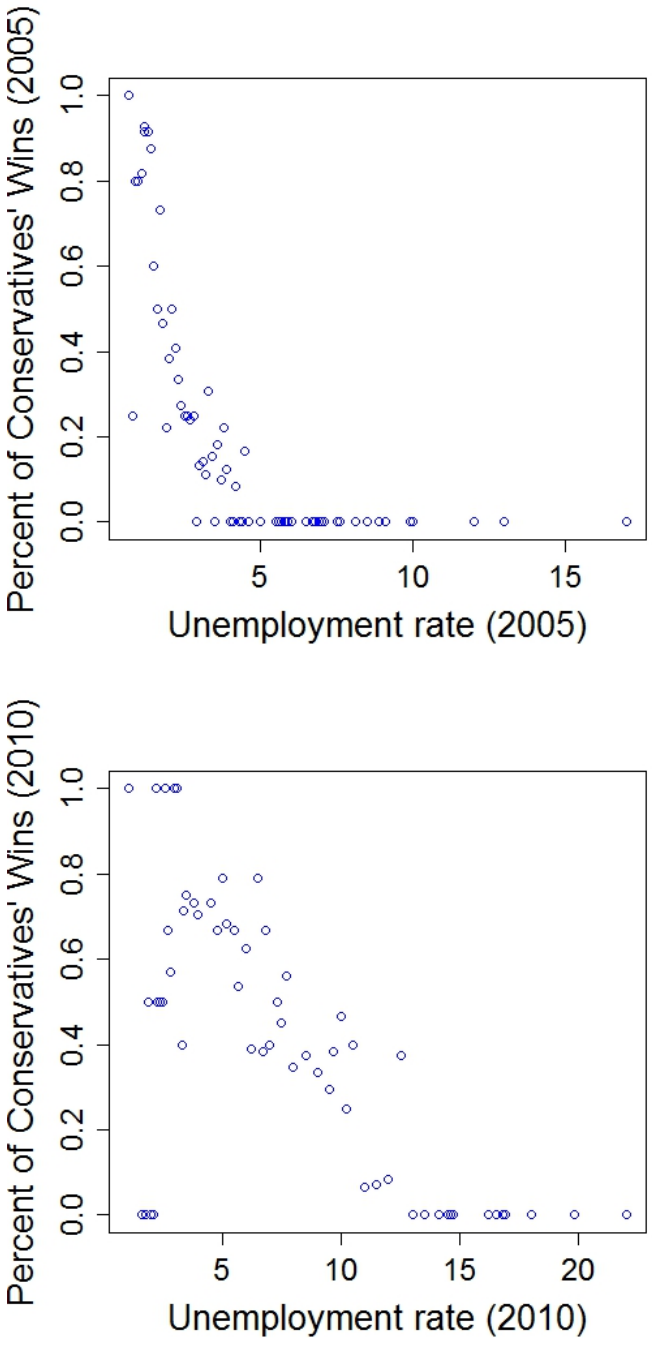


FIGURE 5.3: The correlation of Conservatives' wins and unemployment rate for the general elections of 2005 and 2010 (Kalampokis et al., 2013c)

2015. In Figure 5.4 these models are depicted. In particular, the models associate the probability P(A) the Labour party to win in a specific parliament constituency with the unemployment rate in the same parliament constituency using the logistic function. In the Figure the unemployment rate has been normalised using the *scale*() function in order to take into account the fluctuation of unemployment rate over time.

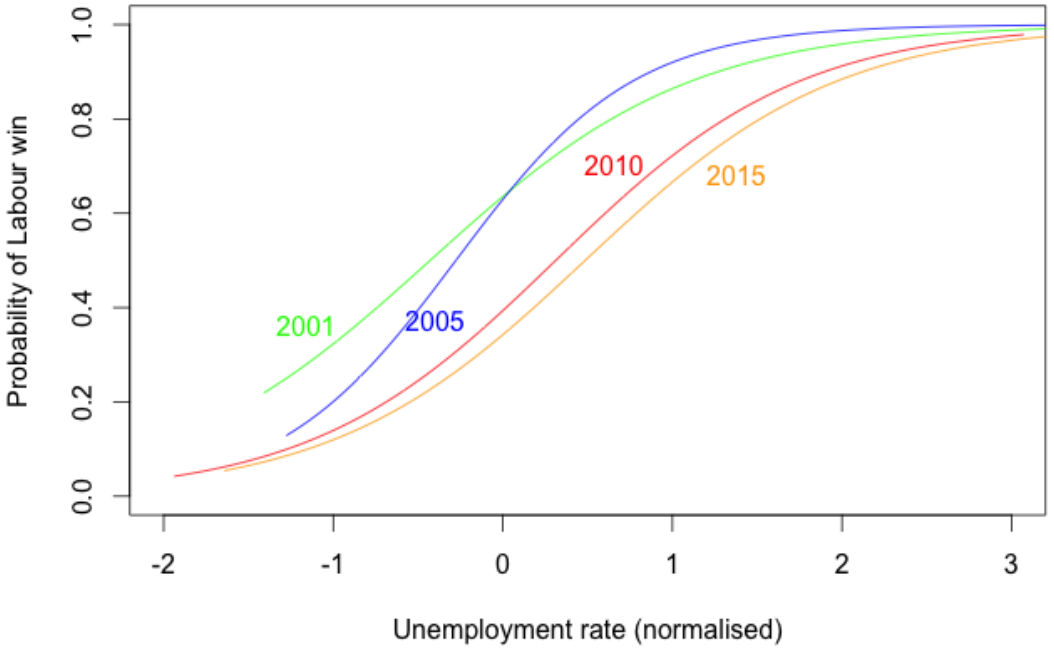


FIGURE 5.4: The probability P(A) the Labour party to win in a specific parliament constituency and the unemployment rate in the same parliament constituency.

Moreover, Table 5.1 presents the Z–statistic and P–value for the four models.

For example in 2010 the probability the Labour party to win in a constituency can be quantified as follows:

$$P(A) = \frac{1}{1 + e^{-y}}$$

where

$$y = -3.823 + 0.437 \times x$$

and x is the unemployment rate of the constituency. So for example, in constituencies with 12 percent unemployment rate the y variable is y=1.421 and the probability P(A)=0.8.

TABLE 5.1: Z-statistic and P-value for the logistic models of UK elections

	Intercept		Unemployment	
	Z–statistic	*P–value*	*Z–statistic*	*P–value*
2001	5.79	7.04e-09	10.14	< 2e-16
2005	6.941	3.91e-12	10.446	< 2e-16
2010	-4.505	6.63e-06	11.069	< 2e-16
2015	-6.441	1.19e-10	10.284	< 2e-16

The UK elections case could also demonstrate the applicability of the linked open government analytics approach into a business setting where enterprise's own data could be combined with OGD. In this case the election results could have been replaced by product sales related data enabling this way the better understanding of an enterprise's sales.

5.3 OLAP Analytics

In this section we describe the development of an OLAP browser for linked data cubes. The resulted software tool is the first browser that enables combining datasets from multiple sources and performing OLAP analytics on top of multiple datasets that reside in disparate sources into the linked data Web.

Given the multidimensional view of data, several decision support operations have been proposed. For example, Ravat et al. (2008) described the following core OLAP operators that allow the expression of multidimensional OLAP analyses:

1. Modifying the analysis precision: It includes (a) selecting data of a multidimensional schema based on a condition that may be expressed on dimension attribute values as well as on fact measure values (slice or dice in a multidimensional databases terminology) and (b) moving the analysis details along a hierarchy (drill-down or roll-up).

2. Changing analysis criteria: It includes (a) replacing a dimension by another one, (b) modifying the analyzed measure set, (c) transforming an analysis axis by

adding or removing a dimension attribute, and (d) adding attributes from external dimensions in a displayed analysis axis.

3. Changing the multidimensional table presentation: It includes (a) switching parameter values of a displayed dimension and (b) add totals and subtotals in a table (this realizes the Cube operator defined by Gray et al. (1997)).

The need for performing OLAP operation on top of graphs (Chen et al., 2009; Zhao et al., 2011) and RDF data (Nebot and Berlanga, 2012) has been identified and introduced in the literature. Moreover, a formal framework for warehouse-style RDF analytics has been introduced and implemented in a platform (Colazzo et al., 2014). This framework employs entity-based RDF data and not aggregated multidimensional cube.

In some cases, a middleware between an RDF application and a typical OLAP server that manages multidimensional data is used (Ghasemi et al., 2014; Kämpgen, 2011). This middleware typically employs a mapping language to convert linked data cubes into a format suitable for loading into an OLAP system. As a result, standard OLAP platforms can be used to elaborate querying and visualising integrated linked data cubes. However, in later works, OLAP operations were transformed into SPARQL queries against linked data cubes (Kämpgen and Harth, 2013; Kämpgen et al., 2012). In particular, Kämpgen et al. (2012) proposed to transform OLAP operations into SPARQL queries against linked data cubes. The performance of these SPARQL queries that has been mapped to typical OLAP operations has been also evaluated (Kämpgen and Harth, 2013). Beheshti et al. (2012) proposed to express OLAP queries in an extension of SPARQL. Moreover, Etcheverry and Vaisman (2012) studied OLAP analysis on Linked Data cubes, while in a latter work Etcheverry et al. (2014) they presented the *QB4OLAP* RDF vocabulary, which aims at extending the QB vocabulary to better support the multidimensional model and better describe hierarchies.

The development of our OLAP browser is based on the theoretical analysis described in section 4.3 and it follows a two cycles approach. The first release of the browser was evaluated by employees from the Research Centre and the Open Data Team of the Flemish Government in the course of the EU funded OpenCube project[1]. Based on the feedback received the Browser was improved and the final version was produced.

[1] http://opencube-project.eu

At the implementation level the Information Workbench (IWB) platform (Haase et al.) serves as the backbone of the Browser. Although IWB provides access to a local RDF store and enables retrieving data by means of SPARQL queries, the components can also access remote RDF stores on the Web using the SPARQL 1.1 federation capabilities. Virtuoso is used as a backed RDF store instead of the native Sesame RDF repository used by IWB. The user interface design is based on the use of wiki-based templates providing dedicated views for RDF resources.

5.3.1 First release of the Browser

The first release of the Browser supports the following functionalities (Kalampokis et al., 2014):

- It presents in a table the values of a two-dimensional slice of a linked data cube. The user can change the number of rows of the table (by default the browser presents 20 rows per page).
- The user can change the 2 dimensions that define the table of the browser.
- The user can change the values of the fixed dimensions (i.e. the dimensions of the cube that are not shown in the table) and thus select a different slice to be presented.
- The user can remove dimensions of the cube to browse. This functionality is supported only for cubes having at least one aggregatable measure.
- The user can create and store a two-dimensional slice of the cube based on the data presented in the browser.

By default, the Browser defines and presents a two-dimensional slice of the cube in the following way:

- It assumes that all the dimensions of the cube will be included in the browser.
- It selects the largest dimension as rows dimension.
- It randomly selects the columns dimension.

- It sets a fixed value for each of the other dimensions (the first value as it appears).
- It randomly selects one measure (in the case of cubes having multiple measures).

In Figure 5.5 the interface of the Browser is depicted. On the top of the page the user can select the dimensions of the cube to browse. In particular, the check boxes enable the insertion or reduction of dimensions. Below the check boxes the actual table is presented while below the table the drop-down lists enable users to change the dimensions that are presented in the table and the values of the fixed dimensions. Finally, at the bottom of the page the use can create and store a slice as it is presented in the browser.

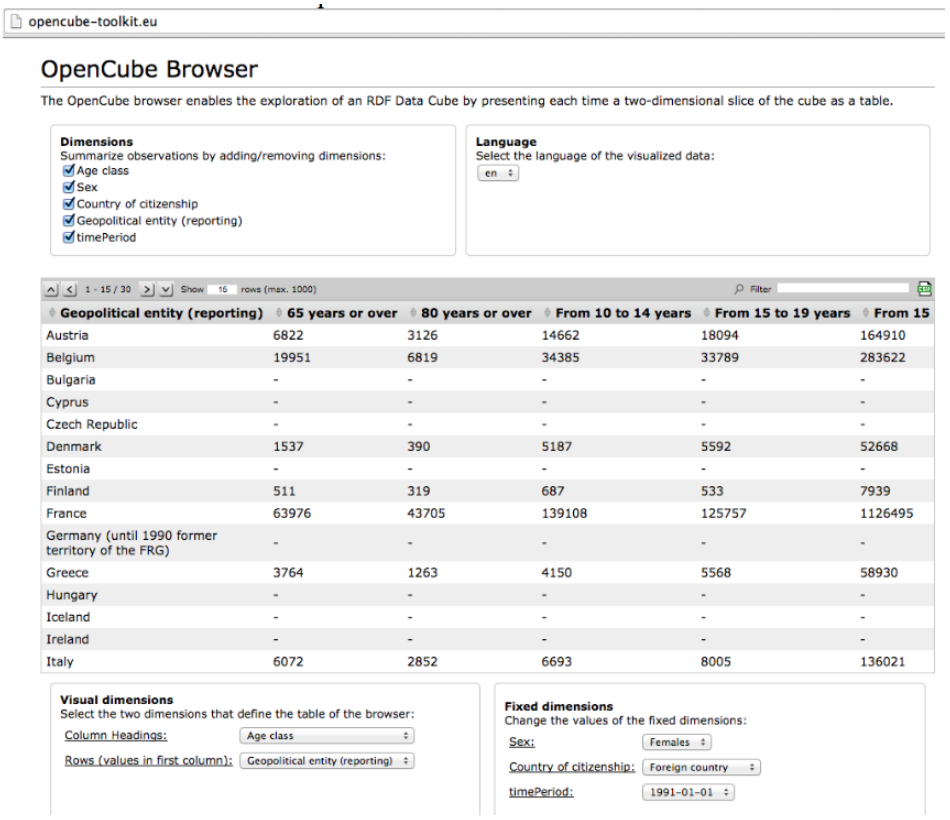


FIGURE 5.5: A linked data cubes browser

Moreover, the functionality provided by the Browser was also implemented on a map of cubes with geospatial dimensions. This geospatial browser enables the visualization of linked data cubes on a map based on their geospatial dimension. In the first release this Browser supports markers, bubbles and choropleth maps. In Figure 5.6 a data cube is visualized on a map based on its geospatial dimension property using a choropleth heat map.

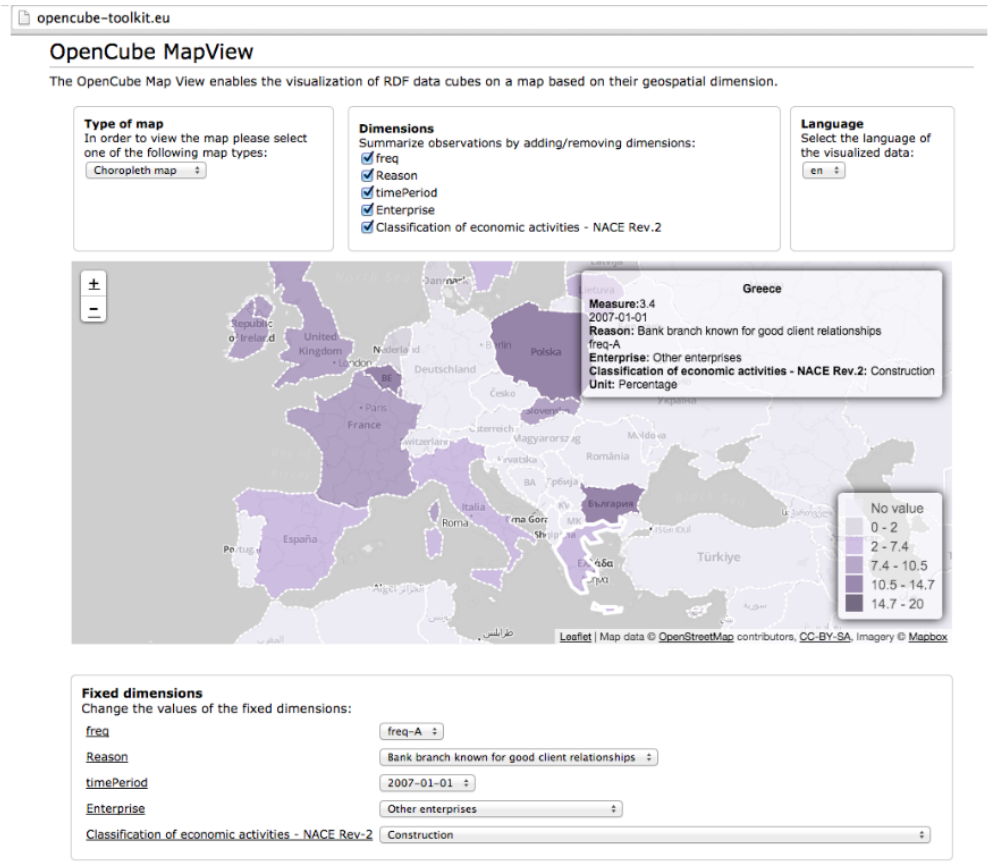


FIGURE 5.6: A tool for browsing linked data cubes on a map

5.3.2 Evaluation of the Browser

In general, the feedback received by the employees of the Flemish Government should be understood in the context of a department seeking to replace an existing solution, which is expensive and not user friendly. Although the overall feedback was positive the following remarks and comments for improvement were expressed (Kalampokis et al., 2016d):

- The multilinguality of the tool was considered as a very important feature.
- Although the performance of the tool was considered acceptable, some users requested better response time.
- The users requested to be able to perform drill-down and roll-up operations over hierarchical code lists (e.g. in geo-spatial dimensions to be able to move across different levels i.e. municipality – district – province – region).

- The users suggested that the interface of the Browser is not clear and easy to use. They proposed to bring all configuration widgets above the table and a dynamically adapted title should describe what is shown.
- The users suggested that the dimension insertion and removal feature is not clear for an average user e.g. a citizen.

Moreover, we should note that the employees of the Flemish Government evaluated Browser in relation to several demos of relevant tools. In this context, their attitude towards the tool is best summarized with a quote from an evaluation form: "We don't see added value compared to other tools". The rationale was that, for the moment, the promise of providing added value through linked data integration across the Web was not visible yet.

Summarizing, the main points expressed in the first phase of evaluation are the following (Kalampokis et al., 2016d):

- The performance of the tools needs to be enhanced.
- Much more attention should be drawn to usability.
- OLAP operations should be enabled in the next phase of the Browser.
- Data cubes integration should be available in a transparent to the user manner.

5.3.3 Second release of the Browser

In this sub-section we describe the second release of the Browser that enables performing OLAP analytics on top of multiple cubes on the Web that reside in disparate sources. The platform is based on and implements the theoretical framework presented at Chapter 4. In particular, the platform comprises the following components: (a) the Aggregator, (b) the Compatibility Explorer, (c) the Expander, and (d) the OLAP browser, which are described in the rest of this section.

Figure 5.7 depicts the architecture of the proposed platform and describes how the components interact.

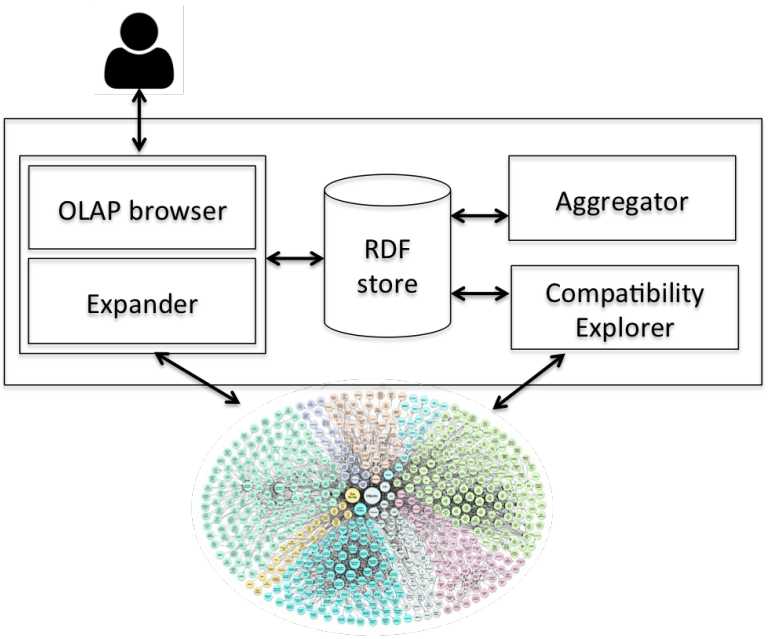


FIGURE 5.7: Technical platform for enabling enhanced analytics on the Web of Linked Data

5.3.3.1 Aggregator

The Aggregator computes aggregations of cells across dimensions or hierarchies. Its role is twofold. First, given an initial cube with n dimensions the aggregator creates $2^n - 1$ new cubes taking into account all the possible combinations of the n dimensions. Second, given a cube and a hierarchy of a dimension, the aggregator computes values of cells for all attributes of the hierarchy that are at a higher level than the attribute of the original cube. Towards this end, it supports three types of aggregate functions as they have distinguished in the literature: Σ, applicable to data that can be added together, ϕ, applicable to data that can be used for average calculations, and c, applicable to data that is constant, i.e., it can only be counted. Considering only the standard SQL aggregation functions, we have that Σ = {SUM, COUNT, AVG, MIN ,MAX}, ϕ = {COUNT, AVG, MIN, MAX} and c = {COUNT}.

For example, the Aggregator computes the SUM of sales of all products of a company for all years and stores and thus creates a new cube having only time and stores dimensions. Moreover, if a cube contains sales per product, day, and store, the Aggregator can compute the sales per product and store for months, quarters, and years.

Finally, in the case of new cubes creation we introduce the concept of *Aggregation Set* to denote the set of cubes created by computing aggregations across all dimension

combinations of an initial cube. As a result, we define the class $ext : AggregationSet$ and the property $ext : aggregationSet$. The Aggregator creates an instance of this class for every set created by an initial cube and connect every cube of the set to this instance using the $ext : aggregationSet$ property. We use the $ext :$ namespace to denote that these classes and properties are potential extensions of the QB vocabulary.

Finally, we should note that the user specifies the aggregation function that will be used in a particular measure and/or dimension. This process, however, can be automated in the case that relevant metadata are added to the description of the cube.

5.3.3.2 Compatibility Explorer

Given a cube in the local RDF store of the platform, the main role of the Compatibility Explorer is to (a) search into the Linked Data Web and identify cubes that are compatible to expand the initial cube, and (b) establish typed links between the local cube and the compatible ones.

Compatibility explorer checks whether a cube in a remote store fulfils the requirements for the four cases of compatibility as described in sub-section 4.1. Towards this end, the following mappings between the proposed algebra and linked data concepts are considered:

- Equality between two dimensions is checked at the data structure definition (DSD) using (a) URIs of the dimensions or (b) URIs of the code lists that are connected to the dimension through the $qb : codeList$ property.
- Equality between two measures is checked at the DSD using URIs of measures.
- Equality between two objects is checked at the observation level using URIs of objects. Objects are expected to be defined as $skos : Concept$.
- Equality between two attributes is checked at observation level using URIs of attributes. Attributes are expected to be defined as $xkos : ClassificationLevel$.

If a compatible cube is identified a typed link is established between the two cubes. Towards this end, we introduce one class and three properties that may extend QB vocabulary as presented in Figure 5.8. We use the $ext :$ prefix to denote these new class and

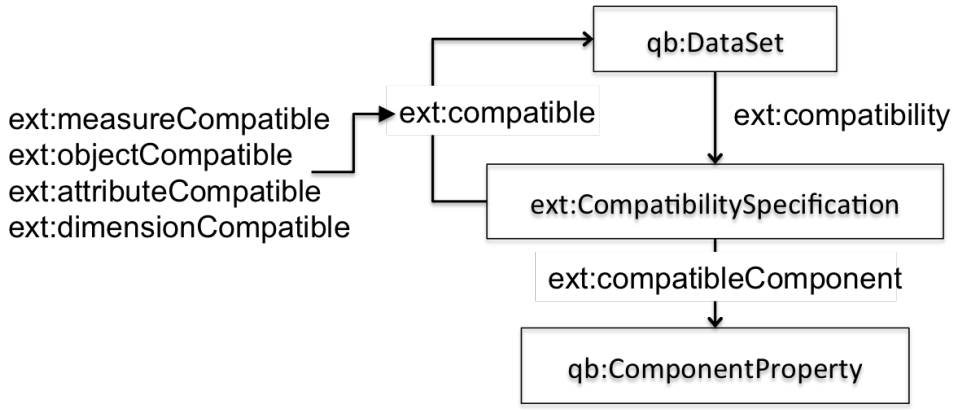


FIGURE 5.8: Proposed extension to the QB vocabulary

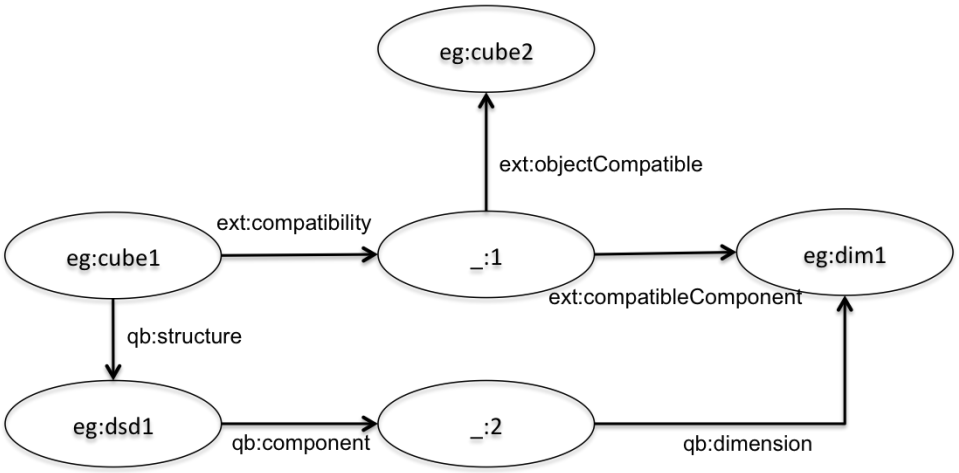


FIGURE 5.9: An example of linking two compatible linked data cubes

properties. To specify the link one needs to declare a $ext : CompatibilitySpecification$ resource which in turn will reference to the identified compatible cube (denoted as a $qb : DataSet$ resource) through the $ext : compatibility$ property. In some cases, however, two cubes are compatible only for a particular $qb : ComponentProperty$. Consider the case where a cube is compatible to expand an original cube by adding objects to a particular dimension. In that case the $ext : CompatibilitySpecification$ resource should reference to this dimension. An example of linking two compatible cubes is presented in Figure 5.9. The $eg :$ prefix denotes an example prefix.

The explorer checks all four cases of compatibility as defined in section 4, and thus four sub-properties of $ext : compatible$ property are introduced as depicted in Figure 5.8

5.3.3.3 Expander

The Expander creates a new expanded cube by merging two compatible ones based on the operators described in sub-section 4.2. In order to discover the compatible cubes the Expanders employs the links established by the Compatibility Explorer.

Initially, the Expander presents to the user cubes that are connected to at least one other cube meaning that it presents only expandable cubes. The user selects a cube and one of the four possible ways that are available to expand it. Thereafter, the Expander presents all the compatible cubes that can expand the cube at hand based on the specific type of compatibility. The user selects one out of the available options and a new linked data cube can be created and stored in the local store.

The Expander, however, can be also integrated with another component that performs some sort of statistical analysis on the data. In this case, it does not create and store a new data cube but an integrated view of the two cubes, which can be further analysed. In the platform that we present at this article the Expander is integrated with an OLAP browser enabling this way the performance of OLAP operations on top of integrated views of compatible cubes on the fly without requiring the actual creation and storage of the expanded cube.

Let us consider a case where a user starts with a cube from the World Bank with two dimensions, namely *time* and *countries*, and one measure, namely *proportion of seats held by women in national parliaments*. Using the Expander the user selects a cube from UNESCO in order to add a new measure into the initial cube. The interface of the Expander is presented in Figure 5.10.

5.3.3.4 Linked Data OLAP Browser

The linked data OLAP browser exploits the others components of the platform in order to enable performing OLAP operation on top of expanded cubes that include measures, dimensions, objects, and/or attributes from multiple cubes that reside on disparate sources on the Web.

Initially, the browser starts with an empty canvas and presents only the structure of the expanded cube (available dimensions and measures), while the user has to select

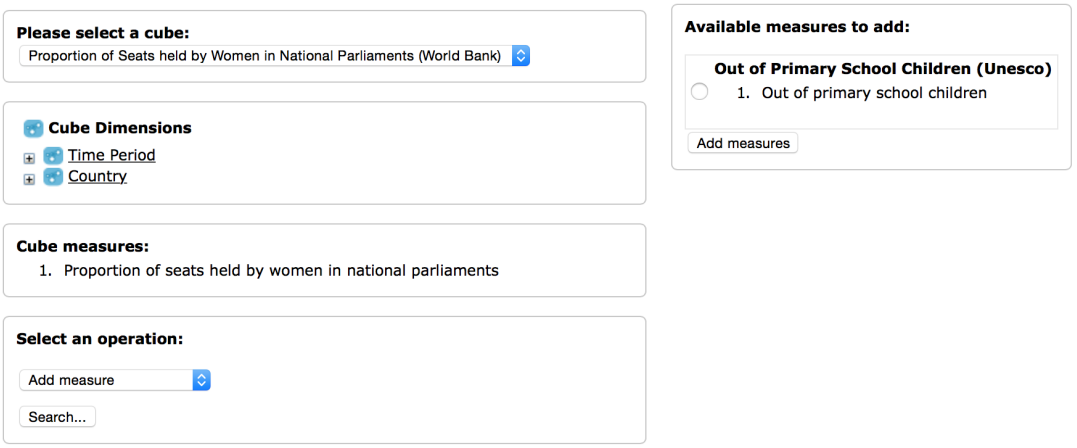


FIGURE 5.10: The Expander

at least one dimension and one measure to visualise. In the case that a dimension has multiple attributes then the browser also presents these attributes and the user has to select at least one of the attributes. Figure 5.11 presents the interface of the browser. In particular, the browser enables a user:

- Exploring a linked data cube by presenting two-dimensional slices of the cube in a table.
- Changing the axes of the table and also changing the object of the fixed dimensions and thus select a different slice to be presented.
- Adding or removing dimensions to be presented in the table.
- Adding or removing measures to be presented in the table.
- Performing roll-up and drill-down across hierarchies.
- Changing the language. This requires the use of multilingual $skos : labels$ available in a dataset.

The browser sends a SPARQL query to retrieve the dimensions, attributes, objects, and measures. Based on the selections of the user, the browser sends a query to retrieve the values of the measures for each cell. The dimensions and measures are retrieved from the metadata of the cube (i.e. from the $qb : DataStructureDefinition$), whereas the attributes, objects, and values of the measures from the actual data (i.e. from the $qb : Observations$).

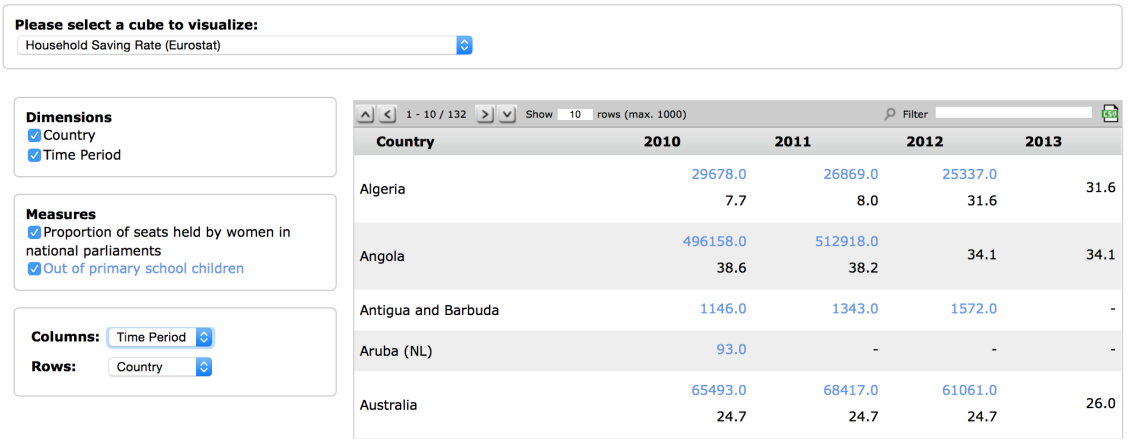


FIGURE 5.11: The linked data OLAP browser

When a user selects to add or remove a dimension, the browser exploits the links established by the Aggregator. In particular, it identifies and presents a cube that has the new set of dimensions and belongs to the same *AggregationSet* as the initial one. On the other hand, when a user wants to move across a hierarchy the browser exploits the hierarchical code lists that have been defined with SKOS and XKOS.

In the case of browsing an integrated view of two cubes with more than one measure then the browser presents the values of these measures using different colours.

5.3.4 The case of Flemish Government

An instance of the developed platform have been deployed at the premises of the Flemish government. Flemish government had already opened up statistics by means of linked data cubes. In particular, 11 cubes had been transformed to linked data according to the QB vocabulary and stored in a Virtuoso RDF store. The deployed platform has been connected to this store.The provided cubes include data regarding unemployment, social housing, real estate prices, etc. and their most common dimensions are time, geography, age group, and sex. The geospatial dimension comprises four attributes, namely region, province, district, and municipality, while the age group dimension comprises three levels:

- level 1: 0-24, 25-49, 50+,
- level 2: 16-24, 25-34, 35-49, 50-64, 65+,
- level 3: 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64

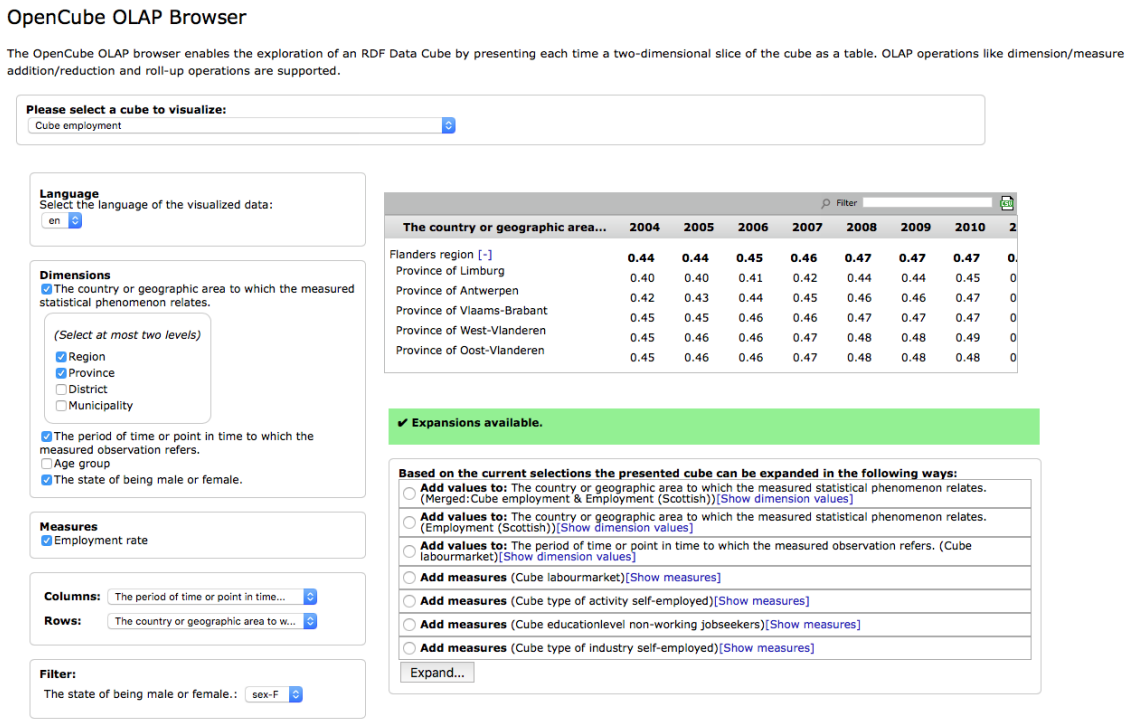


FIGURE 5.12: OLAP Browser as deployed at Flemish Government

Using the components of the platform a total of 186 new aggregated cubes have been created and stored. Moreover, for all cubes that had measured values at the lowest level of a hierarchy new observations have been computed for all the higher levels. Finally, 388 links have been established from 82 cubes or sub-cubes created by the Aggregator to other compatibles cubes or sub-cubes. Moreover, links have been established to three linked data cubes that are published by the Scottish Government. At this point, we need to recall the analysis presented in section 4.3.2.3 on the challenges related to the integration of data cubes coming from different sources. These challenges hamper data integration and thus the exploitation of combined data becomes very difficult.

A user can access and browse all the available cubes using the OLAP browser in the following URL: `http://188.166.18.242:50080/resource/OpenCubeOLAPBrowser`. Figure 5.12 presents the OLAP browser as used at the Flemish government presenting a cube along with the available expansions that can be applied. At the left hand side the structure of the cube (i.e., measures and dimensions along with their levels) is presented. The user can select which of these will be included in the table that is presented at the right hand side of the interface. The table presents the actual values of the cells based on the selections of the user. If any links to other cubes are available based on the current selections, then the Browser presents under the table the expansions that could

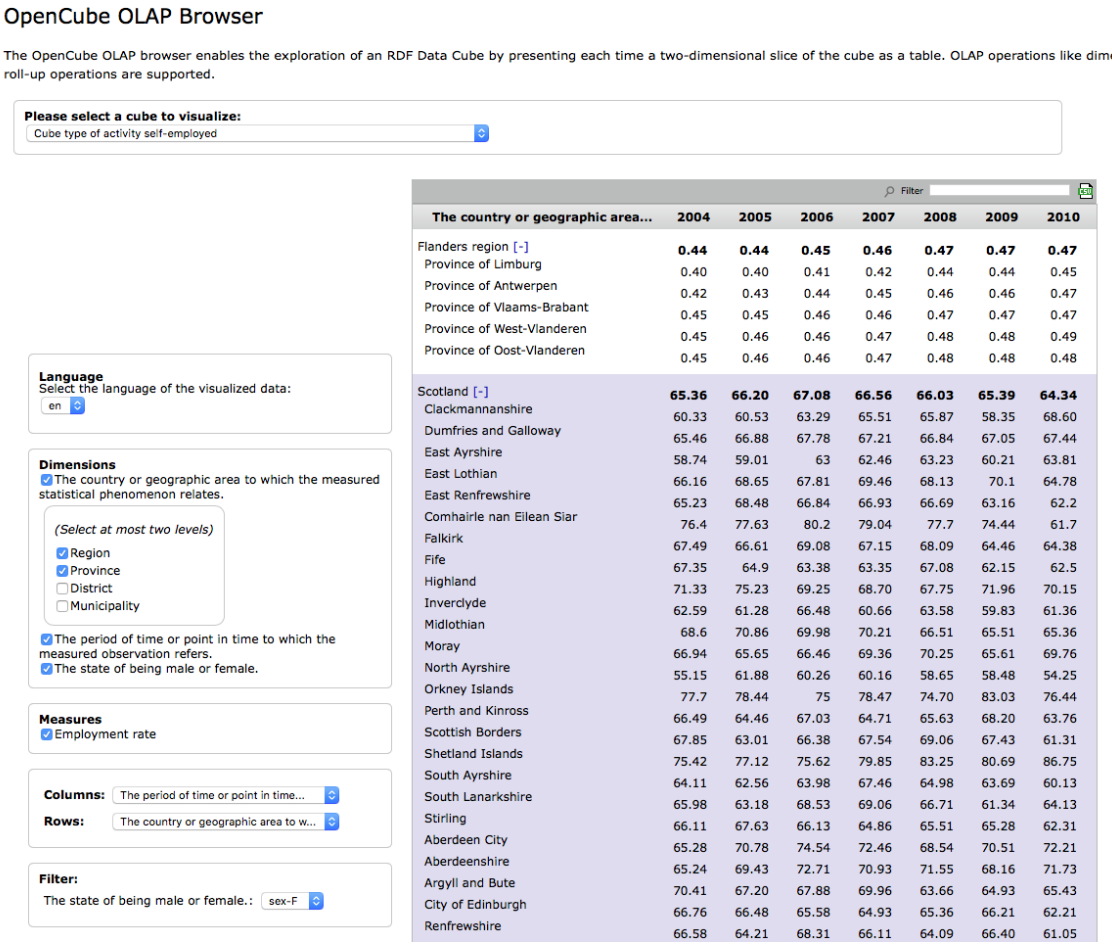


FIGURE 5.13: OLAP Browser as deployed at Flemish Government

apply. In the example presented in Figure 5.12 seven different options for expansion are presented. Three out of the seven options will add new members on the dimensions while the other four will add new measures to the initial cube. Moreover, the first two options will bring data from the Scottish Government. As a result, in the case that a user selects to add new members to the geospatial dimension from Scotland then a new integrated view of the cubes is created. In Figure 5.13 this new expanded view that integrates one cube from Flemish and one from Scottish government is presented. The drill-down and roll-up functionalities are also presented in the Figure as the user can move from the Region level to the Province level in both countries. Finally, the user can change the dimensions that are presented in the table. In the specific example, the gender dimension has been also selected but it has been fixed in the 'Female' value.

5.4 Predictive Analytics

In this section the use of Open Data in predictive analytics is explored. We start from Social Media (SM) data and explore whether or not this type of data is reliable source of input in predictive analysis studies. Thereafter, we exploit the results of this step in an experiment where SM data are combined with other linked data on the Web in order to predict the result of the UK elections of 2010.

5.4.1 Social Media Data in Predictive Analytics

Social Media (SM) data incorporates personal opinions, thoughts and behaviours making it a vital component of the Web and a fertile ground for a variety of business and research endeavours. In this context, the predictive power of SM has been recently explored. For instance, empirical studies have analyzed the Yahoo! Finance message board to predict stock market volatility (Antweiler and Frank, 2004), weblog content to predict movies success (Mishne and Glance, 2006), Google search queries to track influenza-like illnesses (Ginsberg et al., 2009), Amazon reviews to predict product sales (Ghose and Ipeirotis, 2011) and Twitter posts (aka tweets) to infer levels of rainfall (Lampos and Cristianini, 2012).

These research efforts require cross-disciplinary skills as they involve both the transformation of noisy raw SM data into high quality data suitable for statistical analysis as well as the creation of statistical models to estimate the actual outcome. In this setting, a number of researchers have recently challenged the methods employed and the results reported by empirical studies in the area. For instance, Jungherr et al. (2012) repeated the study conducted by Tumasjan et al. (2010) and reported controversial results. In addition, Gayo-Avello (2011) and Metaxas et al. (2011) conducted a number of experiments and criticized generalizations regarding the predictive power of SM.

This chapter aims at consolidating the knowledge created by empirical studies in recent years that exploit SM for predictions, thus enabling an in-depth understanding of SM predictive power. More specific objectives are: (a) to identify steps that characterize all relevant studies as well as approaches that can be followed in each step, and (b) to understand how different steps and approaches are related to SM predictive power.

In order to achieve our objectives we capitalize on the method proposed by Webster and Watson (2002) for conducting systematic literature reviews in the field of information systems. Initially, we performed a systematic search in order to accumulate a relatively complete body of relevant scientific literature. Towards this end, we started with Google Scholar using the key words predict OR forecast AND social media and we collected an initial pool of articles. Thereafter, we went backward by reviewing citations in the identified articles and forward by using Google Scholar's functionality to identify articles citing the previously identified articles. We thereafter studied and filtered these initially identified articles in order to come up with the final set that was included in our research. For this purpose, we used the following inclusion and exclusion criteria:

- We excluded qualitative or purely theoretical articles (e.g. Louis and Zorlu, 2012).
- We included only studies aiming at making predictions. As a result, we have excluded empirical studies that aim at studying the relationship between SM data and phenomena outcome following an explanatory approach (e.g. Chen et al., 2011; Chevalier and Mayzlin, 2006; Chunara et al., 2012; Corley et al., 2010; Duan et al., 2008; Morales-Arroyo and Pandey, 2010; Reinstein and Snyder, 2005; Ye et al., 2009).
- We included only studies that attempt to predict real world outcomes. Thus, we excluded studies that predict online features such as tie strength (Gilbert and Karahalios, 2009), volume of comments on online news (Tsagkias et al., 2010) or movie rating on IMDB (Oghina et al., 2012).

This approach resulted in a set of 52 articles.

In order to synthesize the accumulated knowledge we performed a concept-centric analysis. The main steps and most important aspects composing the whole prediction analysis process were extracted and combined in a conceptual SM data analysis framework for predictions that structures and depicts the area. Finally, the framework was employed to further analyze the literature and to extract insights into the predictive power of SM.

5.4.1.1 A framework for social media data analytics

The proposed framework comprises two discrete phases, namely the *Data Conditioning Phase* and the *Predictive Analysis Phase*. The former refers to the transformation of noisy raw Social Media (SM) data into high quality data that is structured based on some predictor variables. The latter phase refers to the creation and evaluation of a predictive model that enables estimating outcome from a new set of observations.

Each of these phases can be further divided into a sequence of stages and each stage into a number of steps. Finally, different approaches can be followed in each step. Figure 1 presents our framework with the two phases, the respective stages along with their steps.

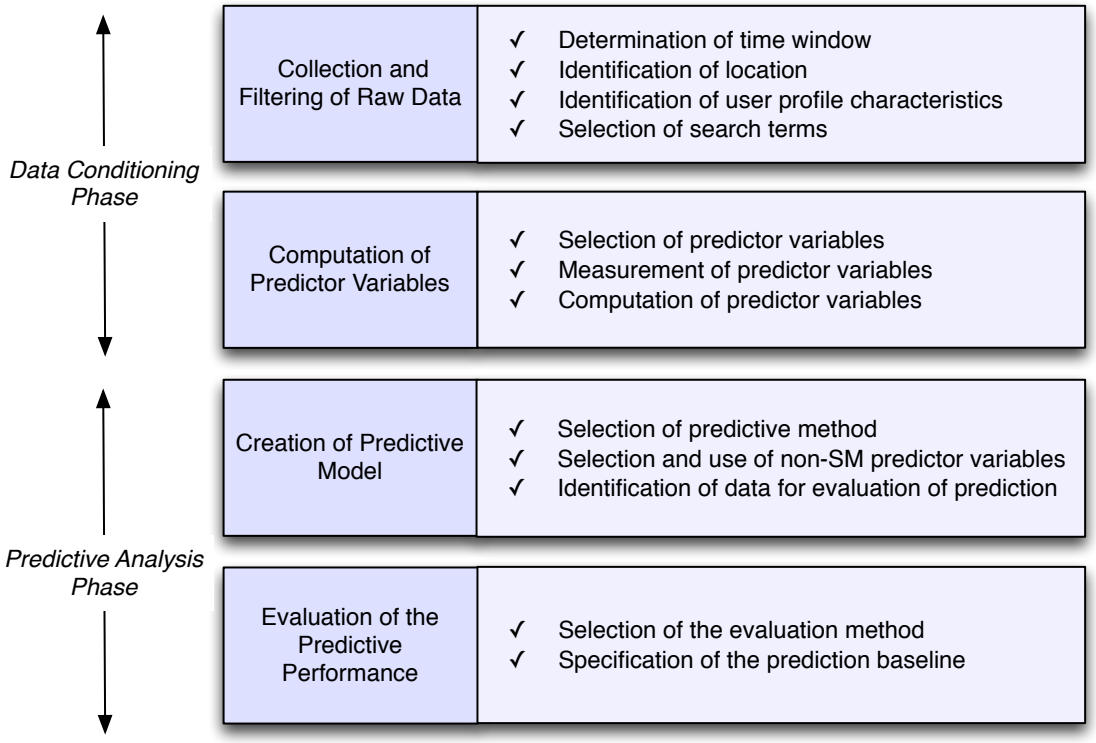


FIGURE 5.14: The Social Media Data Analytics Framework (Kalampokis et al., 2013b)

Phase 1: Data Conditioning

The main purpose of the Data Conditioning phase is the transformation of noisy raw SM data into high quality data that will enable the computation of predictor variables. In order to define data quality we adopt and adapt a model proposed by Strong et al. (1997). In particular, we employ three data quality dimensions from Strong's model that

we consider important in the SM data analysis realm, namely objectivity, completeness and amount of data.

Data objectivity is related to the accuracy of data production or the accuracy of the interpretation process, and specifies whether data is what it claims to be and measures what is supposed to measure. For instance, the data produced by interpreting text's sentiment could be of questionable objectivity in the case of non-rigorous sentiment analysis. The same holds when irrelevant data is interpreted as relevant. Data completeness deals with missing values from a data analysis perspective. It specifies whether or not collected data cover all aspects of a phenomenon in terms of e.g. entities characterizing it and/or predictor variables. Finally, amount of data (or sufficiency) specifies whether or not collected data is sufficient for predictive analysis.

The stages included in this phase along with the steps in each stage are described below.

Stage 1.1: Collection and Filtering of Raw Data

This stage deals with both raw SM data collection from various sources and filtering of data in order to determine those relevant. After its completion the final data set that will be further analyzed during the next stage, is produced. In order to determine the relevant raw data, the when, where, who and what questions should be answered. For example, it can be inferred that a tweet mentioning the Conservative Party one week before the UK elections of 2010 is related to these elections. The same holds for a tweet posted by David Cameron in the same period. The information used to determine relevance is extracted from the actual SM data or their metadata.

The effort required for this stage depends on both the SM and the application area. For example, data filtering in Twitter is challenging because of its noisy nature, while in Amazon it is straightforward as the reviews are aggregated in the product's Web page. Detailed steps that are involved in this stage are described below.

Determination of time window

The time window is related to the when question as it specifies the duration of the collection activity as well as its relation to the characteristic period of the phenomenon. The characteristic period for product sales could be related to the new-product lifecycle (Liu et al., 2010), while for a disease outbreak to duration of pandemic stages (Ritterman

et al., 2009). Clearly, the time window affects both the completeness and the sufficiency of the data.

Identification of location

The identification of location characterizing data is related to the where question. It is crucial in some phenomena (e.g. determination of natural phenomena occurrence) and thus accurate extraction of location is very important. The location characterizing SM data can be extracted from metadata (e.g. Achrekar et al., 2011; Lampos and Cristianini, 2012) or inferred from actual data.

Identification of user profile characteristics

The information related to the online profile of a user answers the who question. In a number of empirical studies (e.g. Forman et al., 2008; Skoric et al., 2012) it is suggested that this information is very important. For instance, Achrekar et al. (2011) filter tweets from the same user within a certain syndrome elapsed time in order to avoid duplication from multiple encounters associated with a single episode of the illness.

Selection of search terms

The search terms selection step deals with the what question. In complex phenomena the identification of both the complete and correct set of search terms can be challenging. For example, Da et al. (2011) measured the search volume for 3,606 stocks through Google trends based on both stock ticker and company name and they, interestingly, identified that their correlation was only 9 percent. The inadequate completion of this task could result in poor quality data regarding its completeness and objectivity. The different approaches for this step can fall into two broad categories: (a) manual approaches where researchers set search terms (e.g. Polgreen et al., 2008) and (b) dynamic approaches where search terms are derived through a computational process (e.g. Ginsberg et al., 2009). We should note that we consider the use of Google Trends as a dynamic selection approach since the resulting categories are determined based on Google's natural language classification engine.

Stage 1.2: Computation of Predictor Variables

This stage deals with analysis of the raw data resulting from the previous stage in order to compute the values of predictor variables. In this stage, only variables related to

SM are considered despite the fact that more variables (e.g. product price) can be finally employed in the predictive analysis stage. The steps composing this stage are the following:

Selection of predictor variables

Although a number of different variables have been used in the literature, we classify them into the following categories:

- Volume-related variables: these measure the amount of SM data in the form of number of tweets, number of reviews, number of queries etc.
- Sentiment-related variables: these measure the sentiment expressed through the data. The sentiment has been measured in the literature with the bullishness index (Oh and Sheng, 2011), review valence (Forman et al., 2008), review rating (Ghose and Ipeirotis, 2011), etc.
- Profile characteristics of online users such as Facebook friends (Franch, 2013), number of followers of users that posted a tweet (Rui and Whinston, 2012), total posts (Oh and Sheng, 2011), the location of the reviewer (Forman et al., 2008) and in-degree (Livne et al., 2011).

The proper selection of the variables that are employed in the analysis can influence the completeness of the data.

Measurement of predictor variables

The majority of variables are usually measured at successive time instants separated by uniform time intervals and are thus expressed as time series. The time intervals that have been used in the literature vary from hours to months. However, in some cases variables are measured just once hence resulting in one value per variable (e.g. Tumasjan et al., 2010).

Careful selection of measurement time intervals allows predictor variables to be comparable to the actual outcome. For instance, Forman et al. (2008) aggregated data by month because the evaluation data of the outcome was formed in monthly reports. However, in some cases the measurement of variables follows different time intervals than the actual outcome data (e.g. Tumasjan et al., 2010).

Computation of predictor variables

Although the computation of volume-related variables is straightforward and provides accurate results, the computation of sentiment expressed in text can be cumbersome and may provide poor results. Literature reveals that many research efforts have come up with poor sentiment analysis results (e.g. Gayo-Avello, 2011; Metaxas et al., 2011), mainly because of the informal and noisy nature of SM that creates problems to widely used NLP tools. The poor performance of sentiment analysis is a major source of weakness in the quality of data objectivity as the interpreted sentiment is different than that actually expressed. In general the approaches used for sentiment computation can be categorized as follows: (a) lexicon-based, where sentiment is defined by the occurrence in the text of words included in a pre-defined lexicon (e.g. Metaxas et al., 2011; O'Connor et al., 2010) and (b) machine learning, where sentiment is computed by language model classifiers (e.g. Asur and Huberman, 2010).

Phase2: Predictive Analytics

The aim of this phase is the creation and evaluation of a predictive model that will enable accurate prediction of phenomenon outcomes based on a new set of observations, where new can be interpreted as observations in future or observations that were not included in the original data sample.

Statisticians recognize that analyses aimed at prediction are different from those aimed at explanation (Konishi and Kitagawa, 2008). Predictive power refers to the ability of predicting new observations accurately. In contrast, explanatory power refers to the strength of association indicated by a statistical model. Although statistically significant effects or relationships do not guarantee high predictive power, empirical studies that make predictive claims often infer predictive power from explanatory power without employing predictive analytics (Shmueli and Koppius, 2011).

Stage 2.1: Creation of Predictive Model

In this stage the actual model is created based on statistical or data mining methods. The steps that compose this stage are described below.

Selection of predictive method

The actual model of the predictive analysis is built based on different statistical or data mining methods. The most common method in literature is linear regression but many others have been also employed such as logistic regression (Livne et al., 2011), Markov models (Gruhl et al., 2005), neural networks (Bollen et al., 2011), support vector machine (Ritterman et al., 2009) and Granger causality (Gilbert and Karahalios, 2010).

Selection and use of non-SM predictor variables

Apart from the predictor variables computed through SM data, other predictor variables are also used in the predictive model. These usually express objective facts, such as past values of phenomenon outcomes and demographics. For instance, Forman et al. (2008) studied the relation between both the average valence of a review and the percentage of reviews disclosing real name or location, and product sales on Amazon. Towards this end, they also employed product price as a control variable in order to reduce the possibility that results reflect differences in average unobserved product quality rather that aspects of the reviews per se. In addition, Rui and Whinston (2012) employed non-SM predictor variables such as budget of a movie or the fact that a movie is a sequel in order to enhance the accuracy of the model and Da et al. (2011) employed the number of news data from the Wall Street Journal in order to predict stock prices.

Identification of data for evaluation of prediction

The data referred to here represent the actual phenomenon outcome. This data is taken from official sources such as governmental documents and Web sites (e.g. Ettredge et al., 2005; Lampos and Cristianini, 2012; Sakaki et al., 2010), other trustworthy Web sites (e.g. Bollen et al., 2011), international organizations (e.g. O'Connor et al., 2010), etc. The accuracy and timely collection of this data is important for the creation of the predictive model.

Stage 2.2: Evaluation of the Predictive Performance

In this stage prediction accuracy is evaluated against the actual outcome. The steps that comprise this stage are described below.

Selection of the evaluation method

The evaluation of predictive performance is very important as it provides the actual result of the study as a whole. In the literature two different approaches are mainly

employed: (a) explanatory analytics and (b) predictive analytics. The former assesses the statistical significance of the model using metrics such as p-values or R^2 (e.g. Asur and Huberman, 2010). The latter usually obtains out-of-sample data to be used for actual evaluation based on metrics such as out-of-sample error rate and statistics such as Predicted Residual Sums of Squares (e.g. Bordino et al., 2012), Root Mean Square Error (e.g. Achrekar et al., 2011), Mean Absolute Percentage Error (e.g. Bollen et al., 2011; Liu et al., 2007) and cross-validation summaries.

In general, the criteria that specify whether a study follows a predictive evaluation method or not are the following (Shmueli and Koppius, 2011):

- Was predictive accuracy based on out-of-sample assessment?
- Was predictive accuracy assessed with adequate predictive measures?

Specification of the prediction baseline

The baseline for prediction is an important element in the literature as it provides an extra metric for evaluating predictive power. The predictive power of an SM data based model is often judged in relation to statistical models fit with traditional data sources (e.g. Goel et al., 2010; Rui and Whinston, 2012) or past values (e.g. Bollen et al., 2011; Ritterman et al., 2009; Wu and Brynjolfsson, 2009). In addition, the results of prediction are sometimes also evaluated against prior models and approaches (e.g. Ghose and Ipeirotis, 2011).

5.4.1.2 Understanding the Predictive Power of Social Media

We now employ our framework in order to gain insight into the predictive power of Social Media (SM). We initially categorize the identified articles based on the application area studied (Table 5.2) and the type of SM employed (Table 5.3).

Table 5.4 presents classification of the literature according to the approach employed for selecting search terms, which is vital in Stage 1.1 of the framework. The table suggests that the vast majority of the studies employs manual selection methods.

In Table 5.5 the studies that involve sentiment analysis are aggregated and categorized according to the method they have employed. Selecting such a method is important

TABLE 5.2: The application areas studied in the literature (adopted from Kalampokis et al. (2013b))

Disease outbreaks	Achrekar et al. (2011); Althouse et al. (2011); Culotta (2010); Ginsberg et al. (2009); Hulth et al. (2009); Polgreen et al. (2008); Ritterman et al. (2009); Signorini et al. (2011); Wilson and Brownstein (2009)
Elections	Franch (2013); Gayo-Avello (2011); He et al. (2012); Jin et al. (2010); Jungherr et al. (2012); Livne et al. (2011); Lui et al. (2011); Metaxas et al. (2011); Sang and Bos (2012); Skoric et al. (2012); Tumasjan et al. (2010, 2012)
Macroeconomics	Choi and Varian (2012); Ettredge et al. (2005); Guzman (2011); O'Connor et al. (2010); Vosen and Schmidt (2011, 2012); Wang et al. (2012); Wu and Brynjolfsson (2009)
Movies	Asur and Huberman (2010); Bothos et al. (2010); Goel et al. (2010); Krauss et al. (2008); Liu et al. (2007, 2010); Mishne and Glance (2006); Rui and Whinston (2012)
Natural phenomena	Earle et al. (2012); Lampos and Cristianini (2012); Sakaki et al. (2010)
Product sales	Choi and Varian (2012); Forman et al. (2008); Ghose and Ipeirotis (2011); Goel et al. (2010); Gruhl et al. (2005); Jin et al. (2010)
Stock market	Antweiler and Frank (2004); Bollen et al. (2011); Bordino et al. (2012); Da et al. (2011); De Choudhury et al. (2008); Gilbert and Karahalios (2010); Oh and Sheng (2011); Zhang et al. (2011, 2012)

in Stage 1.2 of the proposed framework. In this table we do not include studies that express the sentiment as review ratings since its measurement is straightforward.

Based on the criteria employed by Shmueli and Koppius (2011) we also classify (Table 5.6) literature according to the approach used to infer SM predictive power.

Finally, Table 5.7 categorizes literature according to their final outcome with regard to the predictive power of SM. Some studies provide evidence for both outcomes. These are included in both categories.

TABLE 5.3: The social media analyzed in the literature (adopted from Kalampokis et al. (2013b))

Blogs	De Choudhury et al. (2008); Franch (2013); Gilbert and Karahalios (2010); Gruhl et al. (2005); Liu et al. (2007); Mishne and Glance (2006)
Web search	Althouse et al. (2011); Bordino et al. (2012); Choi and Varian (2012); Da et al. (2011); Ettredge et al. (2005); Ginsberg et al. (2009); Goel et al. (2010); Guzman (2011); Hulth et al. (2009); Lui et al. (2011); Polgreen et al. (2008); Vosen and Schmidt (2011, 2012); Wilson and Brownstein (2009); Wu and Brynjolfsson (2009)
Message boards	Antweiler and Frank (2004); Bothos et al. (2010); Krauss et al. (2008); Liu et al. (2010); Oh and Sheng (2011)
Reviews	Bothos et al. (2010); Forman et al. (2008); Ghose and Ipeirotis (2011)
Microblogs (Twitter and Facebook updates)	Achrekar et al. (2011); Asur and Huberman (2010); Bollen et al. (2011); Bothos et al. (2010); Culotta (2010); Earle et al. (2012); Franch (2013); Gayo-Avello (2011); He et al. (2012); Jungherr et al. (2012); Lampos and Cristianini (2012); Livne et al. (2011); Lui et al. (2011); Metaxas et al. (2011); O'Connor et al. (2010); Oh and Sheng (2011); Ritterman et al. (2009); Rui and Whinston (2012); Sakaki et al. (2010); Sang and Bos (2012); Signorini et al. (2011); Skoric et al. (2012); Tumasjan et al. (2010, 2012); Wang et al. (2012); Zhang et al. (2011, 2012)
Social multimedia (YouTube, Flickr)	Franch (2013); Jin et al. (2010)

By synthesizing Tables 5.2–5.7 we can further analyze the empirical studies in the literature and make some interesting observations.

Search term selection Table 5.4 suggests that although dynamic search term selection is used in most application areas (Table 5.2), it only appears in studies that employ Web search and microblog data (Table 5.3). Furthermore, all these studies support SM predictive power (Table 6) based on predictive analytics (Table 5). In the case of manual search term selection when considering the same two SM categories, the percentage of studies that support SM predictive power falls off to fifty percent (50%). Hence, we can

TABLE 5.4: Classification of literature according to the approach for search term selection (adopted from Kalampokis et al. (2013b))

Manual selection	Achrekar et al. (2011); Althouse et al. (2011); Asur and Huberman (2010); Bollen et al. (2011); Bordino et al. (2012); Da et al. (2011); De Choudhury et al. (2008); Ettredge et al. (2005); Franch (2013); Gayo-Avello (2011); Gruhl et al. (2005); Guzman (2011); He et al. (2012); Jungherr et al. (2012); Liu et al. (2011, 2007); Metaxas et al. (2011); Mishne and Glance (2006); O'Connor et al. (2010); Oh and Sheng (2011); Polgreen et al. (2008); Rui and Whinston (2012); Sang and Bos (2012); Signorini et al. (2011); Skoric et al. (2012); Tumasjan et al. (2010); Wilson and Brownstein (2009); Wu and Brynjolfsson (2009); Zhang et al. (2011, 2012)
Dynamic selection	Choi and Varian (2012); Culotta (2010); Ginsberg et al. (2009); Goel et al. (2010); Hulth et al. (2009); Lampos and Cristianini (2012); Ritterman et al. (2009); Sakaki et al. (2010); Vosen and Schmidt (2011); Wang et al. (2012)

TABLE 5.5: Classification of literature according to the text's sentiment analysis approach (adopted from Kalampokis et al. (2013b))

Lexicon–based	Bollen et al. (2011); Gayo-Avello (2011); Liu et al. (2010); Metaxas et al. (2011); O'Connor et al. (2010); Zhang et al. (2011, 2012)
Machine learning	Antweiler and Frank (2004); Asur and Huberman (2010); Bothos et al. (2010); Gayo-Avello (2011); Gilbert and Karahalios (2010); He et al. (2012); Krauss et al. (2008); Liu et al. (2007); Mishne and Glance (2006); Oh and Sheng (2011); **?**

TABLE 5.6: Classification of literature according to the evaluation approach (adopted from Kalampokis et al. (2013b))

Explanatory evaluation	Antweiler and Frank (2004); Asur and Huberman (2010); Bordino et al. (2012); Da et al. (2011); Ettredge et al. (2005); Forman et al. (2008); Gayo-Avello (2011); He et al. (2012); Jin et al. (2010); Jungherr et al. (2012); Krauss et al. (2008); Liu et al. (2011, 2010); Livne et al. (2011); Metaxas et al. (2011); Mishne and Glance (2006); Polgreen et al. (2008); Sang and Bos (2012); Skoric et al. (2012); Tumasjan et al. (2010); Wilson and Brownstein (2009); Zhang et al. (2011, 2012)
Predictive evaluation	Achrekar et al. (2011); Althouse et al. (2011); Bollen et al. (2011); Bothos et al. (2010); Choi and Varian (2012); Culotta (2010); De Choudhury et al. (2008); Franch (2013); Ghose and Ipeirotis (2011); Gilbert and Karahalios (2010); Ginsberg et al. (2009); Goel et al. (2010); Gruhl et al. (2005); Guzman (2011); Hulth et al. (2009); Lampos and Cristianini (2012); Liu et al. (2007); O'Connor et al. (2010); Oh and Sheng (2011); Ritterman et al. (2009); Rui and Whinston (2012); Sakaki et al. (2010); Signorini et al. (2011); Vosen and Schmidt (2011, 2012); Wang et al. (2012); Wu and Brynjolfsson (2009)

conclude that search term selection is of vital importance in microblog and Web search data, and thus these SM categories call for sophisticated search terms selection methods. For instance, Lampos and Cristianini (2012) successfully estimated daily rainfall rates for five UK cities by identifying relevant tweets through the application of Bolasso (i.e. the bootstrapped version of Least Absolute Shrinkage and Selection Operator) for search term selection.

Sentiment analysis Table 5.5 suggests that the majority of studies that employ sentiment analysis investigate stock market and movies (Table 5.2). Although sentiment seems to be important in application areas such as elections, product sales and macroeconomics, only six (6) out of twenty four (24) studies include a sentiment-related independent variable. Disease outbreaks and natural phenomena related studies do not employ sentiment, as one might have expected. Interestingly however, forty per cent (40%) of studies that have used sentiment-related variables challenge SM predictive power. This number increases to sixty five percent (65%) in the case of lexicon-based approaches, while it

TABLE 5.7: Classification of literature based on main outcome (adopted from Kalampokis et al. (2013b))

Support SM predictive power	Achrekar et al. (2011); Althouse et al. (2011); Antweiler and Frank (2004); Asur and Huberman (2010); Bollen et al. (2011); Bordino et al. (2012); Bothos et al. (2010); Choi and Varian (2012); Culotta (2010); Da et al. (2011); De Choudhury et al. (2008); Ettredge et al. (2005); Forman et al. (2008); Franch (2013); Ghose and Ipeirotis (2011); Gilbert and Karahalios (2010); Ginsberg et al. (2009); Goel et al. (2010); Gruhl et al. (2005); Guzman (2011); Hulth et al. (2009); Jin et al. (2010); Krauss et al. (2008); Lampos and Cristianini (2012); Liu et al. (2007, 2010); Livne et al. (2011); Oh and Sheng (2011); Polgreen et al. (2008); Ritterman et al. (2009); Rui and Whinston (2012); Sakaki et al. (2010); Sang and Bos (2012); Signorini et al. (2011); Tumasjan et al. (2010); Vosen and Schmidt (2011, 2012); Wang et al. (2012); Wu and Brynjolfsson (2009); Zhang et al. (2011, 2012)
Challenge SM predictive power	Bollen et al. (2011); Forman et al. (2008); Gayo-Avello (2011); Goel et al. (2010); He et al. (2012); Jungherr et al. (2012); Liu et al. (2011, 2010); Metaxas et al. (2011); Mishne and Glance (2006); O'Connor et al. (2010); Sang and Bos (2012); Skoric et al. (2012); Wilson and Brownstein (2009)

falls off to twenty percent (20%) in those of machine learning. Hence, it seems that sentiment analysis in SM requires innovative approaches that could address the noisy and informal nature of SM.

Evaluation method In general, half of the studies do not use predictive analytics to draw conclusions on the predictive performance of SM. These studies span equally across all SM categories (Table 5.3). With regard to application areas (Table 5.2), the vast majority of election-related cases do not follow a predictive analytics evaluation, while most studies related to macroeconomic indices, natural phenomena and product sales application areas evaluate predictive power based on prediction analytics. The evaluation of a predictive model with out-of-sample data is sometimes challenging. For instance, in the case of election-related studies the outcome is produced once every four or five years. In order to overcome this limitation Franch (2013) used poll data. Tables 5.6 and 5.7 suggest that ten (10) out of fourteen (14) studies that challenge SM predictive power

have used explanatory evaluation methods. This fact does not imply that these studies do not contribute to the understanding of SM predictive power as lack of a statistically significant relationship indicates low predictive power. In addition, fourteen (14) out of forty (40) studies that support SM predictive power infer predictive power without employing predictive analytics. Here we should also note that if these studies had used predictive evaluation methods, they could have presented high predictive power. However, based on the reported results we cannot assess their predictive power because a statistically significant relationship does not always ensure high predictive power. For example, *'low predictive power can result from over-fitting, where an empirical model fits the training data so well that it underperforms in predicting new data'* (Breiman, 2001, p. 204).

Application areas The application area of a study seems to be related to the accuracy of the prediction that the study presents. Some application areas, such as disease outbreak and natural phenomena, do not involve the expression of any kind of opinion or sentiment. The signal that the researcher has to decode in these cases has to do with the occurrence or not of the event. As a result, these studies are expected to provide more accurate predictions than studies requiring extracting opinions or sentiment out of raw data. Moreover, some application areas, such as elections or macroeconomics, can be characterized as complex because they involve multiple and interrelated real-world entities such as political parties and politicians or complex concepts such as consumer confidence or inflation rate. The identification of the complete set of relevant raw SM data in these cases is challenging and hence call for sophisticated methods. This becomes evident if we elaborate on two of the identified applications areas, namely elections and disease outbreak. The former involves opinion expression and is characterized by multiple and interrelated real-world entities (i.e. political parties, candidates, election constituencies), while the latter does not require opinion extraction. Table 5.2 suggests that all eleven (11) election-related studies selected their search terms manually (Table 5.4) and only three of them employed sentiment–related variables (Table 5.5). These facts could provide an explanation of the unfavourable and controversial results reported in the literature regarding predictability of election results through SM. In addition, half of the disease outbreak related studies employed sophisticated search unit selection approaches, eighty percent (80%) used predictive analytics evaluation and ninety percent (90%) supported SM predictive power.

5.4.2 The case of UK Election 2010

So, the proposed framework suggests that all relevant studies can be decomposed into a small number of steps and that different choices can be made in each step. The application of the framework enabled us to make some interesting observations. The majority of the empirical studies support SM predictive power, however more than one-third of these studies infer predictive power without employing predictive analytics. Sophisticated search term selection is crucial in Web search and microblog data. In addition, the use of sentiment-related variables resulted often in controversial outcomes proving that SM data call for sophisticated sentiment analysis approaches.

We anticipate that both the framework and analysis results will enable us to design scientifically rigorous evaluation cases.

5.4.2.1 Predicting Election Results with Social Media

The creation of predictive models using Social Media data has been performed in various application areas such as elections, disease outbreak, natural phenomena, macroeconomics etc. Some of these areas, such as elections or macroeconomics, involve multiple and interrelated real-world entities such as political parties and politicians or complex concepts such as consumer confidence or inflation rate. In this case the identification of the complete and correct set of raw SM data is challenging. Moreover, some of the application areas, such natural phenomena, do not involve the expression of a subjective opinion that may involve sentiment but only the statement of an objective fact e.g. "it is raining". In other areas, however, the expression and thus the computation of the expressed sentiment is crucial. Election related cases can be characterised as complex because they (a) involve multiple and interrelated real-world entities, and (b) require the computation of the expressed sentiment. In this section we, exploit the social media analysis framework to analyse existing research endeavours aiming at predicting elections results through Social Media data Franch (2013); Gayo-Avello (2011); He et al. (2012); Jin et al. (2010); Jungherr et al. (2012); Livne et al. (2011); Lui et al. (2011); Metaxas et al. (2011); Sang and Bos (2012); Skoric et al. (2012); Tumasjan et al. (2010). In particular we concentrate on the steps that were presented in the previous section

and for these steps we identify limitations in existing research endeavors and we state the approach that we follow in this section in order to overcome these limitations.

Search term selection The different approaches for this step can fall into two broad categories: (a) manual approaches where researchers set search terms, and (b) dynamic approaches where search terms are derived through a computational process.

All existing research endeavors use manual approaches to select their search terms that were used to create the pool of data used of their analysis. This is an important restriction taking into account the conclusion of Kalampokis et al. (2013b) that *"search term selection is of vital importance in microblog and web search data, and thus these Social Media categories call for sophisticated search term selection methods"*.

In this section we follow a dynamic search term selection approach by exploiting linked open data (LOD) from DBpedia to enrich tweets. LOD is used to improve the understanding about a tweet and thus facilitate its interpretation. In particular we identify named entities mentioned in the tweets and establish typed links between these entities and representations of the same entities in DBpedia. In order to clarify the proposed approach we describe the case of the United Kingdom (UK) elections. Figure 5.15 depicts four tweets that were published prior to the elections. Each of these includes a different named entity that is related to the elections i.e. Cameron, George Osborn, Tory and Conservatives. In Twitter?s realm these entities and thus the tweets that include them do not have any connection. However, DBpedia has linked data descriptions of the same entities having also links among them. In particular DBpedia suggests that David Cameron and George Osborne belong to the Conservative party and that the Tory party is predecessor of the Conservative party. The integration of the data from DBpedia with the data from Twitter could enable an analyst to understand that the four tweets of figure 5.15 are all refer to the Conservative party.

Selection of predictor variables The variables that are typically used to create predictive models using Social Media data are related to (a) the volume of Social Media posts (e.g. number of tweets, number of reviews etc.), (b) the sentiment expressed through the data, and (c) profile characteristics of online users (e.g. Facebook friends).

Although the majority of the existing research endeavours use volume-related variables in their predictive models, only three of them Gayo-Avello (2011); He et al. (2012); Metaxas

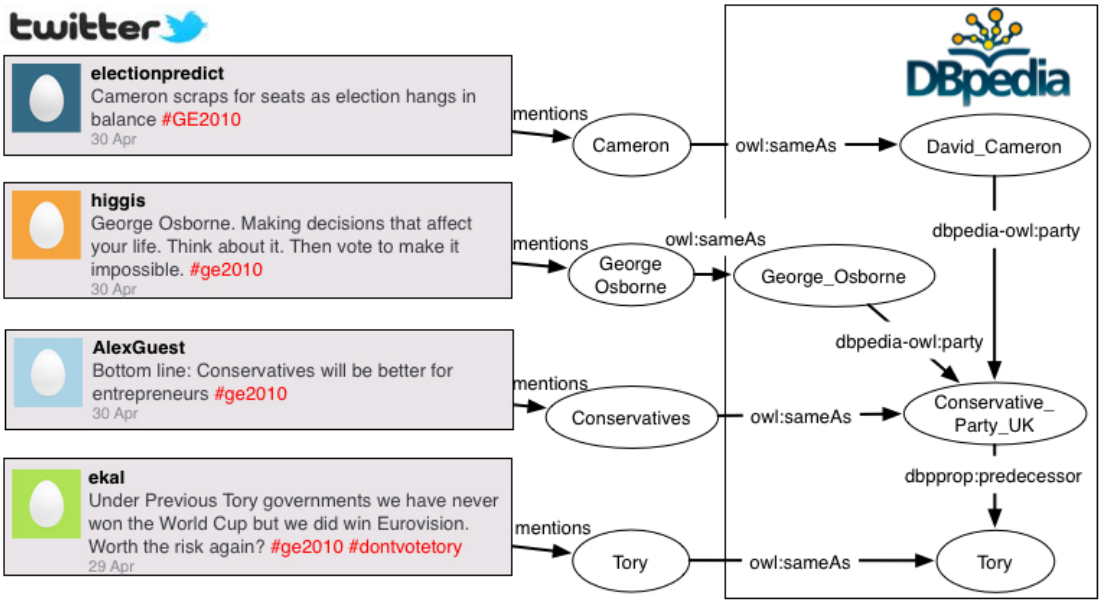


FIGURE 5.15: Linking entities extracted from Tweets to DBpedia

et al. (2011) use sentiment-related variables. This is, however, restricting taking into account the nature of elections where the opinion expressed by the users is important. In this section we exploit both volume and sentiment related variables in order to create the predictive model.

Computation of predictor variables The computation of volume-related variables is straightforward through the count of posts that satisfy some criteria (e.g. the use of a keyword). The computation, however, of sentiment expressed in Social Media may provide poor results and thus impact the prediction accuracy. In general the approaches used for sentiment computation fall into two categories i.e. lexicon-based, where sentiment is defined by the occurrence in the text of keywords included in a pre-defined lexicon, and machine learning, where sentiment is computed by language model classifiers. As Kalampokis et al. (2013b) concluded, the use of machine learning approaches in sentiment analysis supports the creation of more accurate predictive models than the ones created with lexicon-based sentiment analysis.

Out of the 3 studies that employed sentiment related variables in their predictive models, one followed a lexicon-based approach Metaxas et al. (2011), one followed a machine learning approach He et al. (2012), and one used both Gayo-Avello (2011).

In the analysis that we perform in this section we follow a machine-learning algorithm in order to determine the sentiment expressed in tweets and thus compute the value of the variables.

Selection of the evaluation method This step is important because the evaluation of the predictive performance is often not performed using predictive measures and out-of-sample assessment. It is indicative that half o the studies that were reviewed by Kalampokis et al. (2013b) do not use predictive analytics to draw conclusions on the predictive performance of Social Media. When it comes to elections only one study Franch (2013) make use of predictive analytics to predict the results of elections using Social Media data. In this section we use data from polls before the elections in order to create our training data set that is used to develop the model. Thereafter, out-of-sample data are employed to predict the elections results.

5.4.2.2 Description of the Approach

In this section we describe in detail the approach that we followed in each stage of the social media analysis framework in order to predict the results of the UK 2010 elections through the analysis of Twitter data. The election took place on May 6, 2010 in 650 parliament constituencies across the U.K. while Prime Minister Gordon Brown announced the election in April 6, 2010. A total of 49 political parties were participated in the election, with three of them being the most prominent ones i.e. the Conservative party led by David Cameron, the Liberal Democrat party led by Nick Clegg and the Labour party led by Gordon Brown. Smaller parties that finally received more than 1% of the total votes include the UK Independence Party (UKIP), the British National Party (BNP), the Scottish National Party (SNP), the Green Party and the Pirate Party.

Collection and Filtering of Tweets Twitter data was collected between 8/4/2010 and 5/5/2010 (i.e. the last month before the day of the elections) using the Twitter API using the #ge2010, #ukelection, #election2010 and #ge10 hashtags. This resulted in a total of 84.375 unique tweets and 25.241 re-tweets, which we did not remove because we consider them as a way of opinion indication. The retrieved JSON files were stored in an instance of a MongoDB NoSQL database so as to easily access and exploit them.

In order to classify the tweets in the party that they refer to we followed a linked data based approach comprising two steps: i) the extraction of named entities from the text of tweets by employing NER, and ii) the creation of Linked Data including the establishment of links to DBpedia.

In the first step of our approach we used a Conditional Random Field (CRF) sequence classifier provided by Stanford NER . Towards this end, we created the gold standard of our data set by randomly sampling 2000 tweets and manually annotated them with 3 different Named Entity Types: Person, Organization and Location. Thereafter, the classifier was trained on 1000 manually annotated tweets (training data set) from the gold standard and then tested on the remaining 1000 tweets (test data set). The CRF classifier achieved 85.89 F1 score in the data set. The F1 score specifies the accuracy of NER as a harmonic mean of Recall and Precision. Recall is the ratio of the total number of correct entities identified to the total number of entities while Precision the ratio of the total number of correct entities identified tot the total number of entities identified.

During the second step our approach, the extracted entities along with tweet?s metadata were transformed into RDF, which was stored into a Virtuoso store. Thereafter URI aliases were identified between the Twitter data set and DBpedia and owl:sameAs links were established. The aim of this interlinking was to perform a) entity disambiguation i.e. to specify the extracted entities that refer to the same real world entity and b) data enrichment by enabling the use of existing linked data on the Web. In order to perform the interlinking we searched for candidate linking entities in DBpedia using DBpedia Lookup Service . More particularly, we used the Keyword Search API of DBpedia Lookup Service that employs keyword(s) to search for related DBpedia resources. For example, the query `http://lookup.dbpedia.org/api/search.asmx/KeywordSearch?QueryClass=Person&QueryString=Cameron` searches DBpedia for URIs relative to the keyword Cameron. The QueryClass restricts the searching of this query to only resources that are instances of the Person DBpedia class. The specific query returns five results: “David Cameron”, “James Cameron”, “Cameron Diaz”, “Thomas Fairfax, 3rd Lord Fairfax of Cameron” and “Cameron Crowe”.

In our case, DBpedia lookup used the entities identified after the NER process. However, from a total of 6392 distinct entities found in our tweets, DBpedia lookup identified only 3712 matches (58.07%). The unmatched entities include non-real world entities such as

"tory_letter" and entities wrongly annotated by our NER classifier. For example, the classifier annotated William Hague who is a politician as a Location. Thereafter this entity was searched using QueryClass Place in DBpedia lookup, hence returning no results.

In the case where more than one resource was returned by DBpedia lookup, the challenge was to select the most appropriate resource. Towards this end, we initially created a "bag of words" by retrieving and tokenizing the rdfs:label of the resources that are connected with the dcterms:subject property to the returned by DBpedia Lookup resources when searching for the eight most popular UK political parties (i.e. using as keywords "UK_Independence_Party", "Conservative_Party_(UK)", "Labour_Party_(UK)", "Green_Party_(UK)", "British_National_Party", "Liberal_Democrats", "Scottish_National_Party" and "Pirate_Party_UK"). We then created a "bag-of-words" for each of the returning results of each NER entity lookup by retrieving and tokenizing again the rdfs:label of the resources that are connected with the dcterms:subject property to the results . We thereafter computed the TF-IDF vectors for each of the bag of words and compared the TF-IDF vector of each result to the TF-IDF vector of the domain-related bag-of-words using cosine similarity as a measure. The best cosine similarity indicated the most appropriate resource among the resulting resources.

In order to classify the tweets into the different political parties we queried the resulted RDF data in order to identify tweets that have entities that are linked to the DBpedia representation of a political party or to a resource that is connected through the dbpedia-owl:leader, dbpprop:leader, dbpedia-owl:party, dbpprop:party or dbpedia-owl:otherParty properties to a political party. The SPARQL query that we have used is presented in the Appendix D

Computation of Predictor Variables

We selected two predictor variables for our research: (a) relative frequency (RF) of tweets and (b) positive-negative ratio (PNR). The former variable is related to the volume of Social Media data while the latter to sentiment. In particular, we define RF of a political party as the tweet volume per day. For example, if in a particular day a total of 100 tweets that refer UK political parties have been posted and 20 of them regard the Labour party, then the RF of the Labour party at this day will be 0.2 or 20%. On the other hand PNR is used to quantify the sentiment expressed in tweets about a political party.

In particular, for each political party, we define PNR on the day t as the ratio of positive over negative tweets published that day for this political party:

$$PNR_t = \frac{count_t(pos)}{count_t(neg)} \tag{5.1}$$

PNR actually increases when the number of positive tweets for a party is larger than the number of the negative tweets i.e. people tweet in favour of a party.

Our predictor variables were measured using daily time intervals. As a result a number of time series were created to depict the daily measurements of our predictor variables during the whole month before the elections.

We used sentiment analysis to measure our PNR variable. Our sentiment analysis is based on the computation of sentiment by a machine learning language model classifier from the LingPipe package[2]. In particular we employed the DynamicLMClassifier and we performed a k-fold cross-validation to classify the tweets as Positive or Negative. The specific language model classifier accepts training events of categorised character sequences. Training is based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators. Specifically, during a k-fold cross-validation a data set is divided into evenly sized k folds and then k iterations of classifications are performed. Each iteration uses one of the folds of data (a different fold is selected for each iteration) as testing data and the remaining k-1 folds as training data.

In order to train the classifier we need a training data set, which separates positive tweets from negative tweets. Usually the training data set is obtained by manually annotating data but in our case we used a set of positive and negative hashtags (see Table in Appendix D) to identify the relevant categories of tweets. 4053 tweets include negative and 4274 positive hashtags. However, our training data set contains 70% of these tweets because tweets that mentioned more than one political party were ignored so as to be sure that we assign the positive or negative sentiments to the proper political party. We also ignored all re-tweets and, furthermore, preprocessed these tweets in order to:

[2]http://alias-i.com/lingpipe

- Remove user mention entities and URL entities
- Remove phenomenon related hashtags e.g. #ukelection
- Remove all stop-words
- Replace party and candidate names with ¡PP¿ and ¡CA¿ respectively

We thereafter trained the classifier using an n-gram model with n=6. The classifier was trained to predict two classes i.e. positive and negative. Our cross-validation sentiment analysis experiment resulted in a model with an (average) accuracy of 83.67%.

Thereafter we used the classifier to predict the sentiment for all the tweets of each of our two data sets. We applied our sentiment analysis model to all tweets that mention only one political party in order to predict their sentiment. We used a threshold of 85% for the conditional probability of the predicted sentiment in order to keep only high confidence predictions for the sentiment of tweets. The result was 21.487 (out of the 24.265) tweets mentioning only one political party and showing sentiment for it: 10.074 of them showing positive and 11.413 showing negative sentiment respectively.

Creation of Predictive Model

This stage creates the actual predictive models which was based on linear regression analysis. The YouGov data are the dependent variable and RF and PNR the independent variables. We used only one independent variable in these models based on the claim that regressing more independent variables at a time ends up in creating multicollinearity and, therefore, higher standard errors.

The linear regression model involves three variables: i) YouGov data as the dependent variable, ii) RF and iii) PNR having the equation:

$$y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \epsilon \tag{5.2}$$

where y_t is YouGov's measurement of the vote share for a political party on day t, $x_1 t$t is the RF of this political party on day t, $x_2 t$ is the PNR of this political party on day t and ϵ is the error of the regression. The values of b_0, b_1 and b_2 are also computed using least squares analysis.

We have also used smoothing averages over a window of the past k days for the independent variables:

$$MA_t = 1/k(x_{t-k+1} + x_{t-k+2} + \cdots + x_t) \tag{5.3}$$

with k=2, k=3, and k=4.

Evaluation of the Predictive Performance

This stage evaluates prediction accuracy against the actual outcome. For this reason, the models created in the previous stage are initially used to make out-of-sample forecasts. The evaluation of our statistical models is then made using the actual results of the UK elections using predictive analytics as an evaluation method.

Figure 5.16 shows the time series of the YouGov polls regarding the Conservative, Labour, Liberal Democrat and Other parties the last month before the elections. We observe that the general trend for the Conservative, Labour and Other parties lines is to remain almost constant, having few fluctuations. This fact changes however for the Liberal Democrat party where there seems to be a remarkable increase. For this reason, we decided to evaluate the prediction model only for the Liberal Democrat party as making predictions for the rest of the parties will most probably result in extremely accurate results due to the tendency of the YouGov polls to remain almost constant.

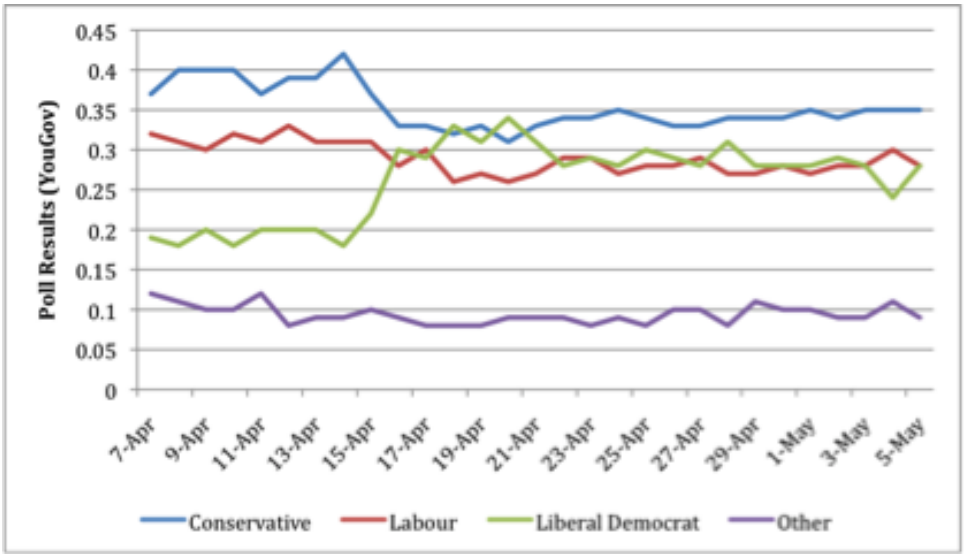


FIGURE 5.16: Time series of YouGov polls over the last month before the election

5.4.2.3 Results

Collection and Filtering of Raw Data

Figure 5.17 depicts the distribution of the total number of tweets regarding all political parties over the whole time period. The figure shows that the majority of the tweets were published during the six last days prior to the election with highest peak the last day before the elections.

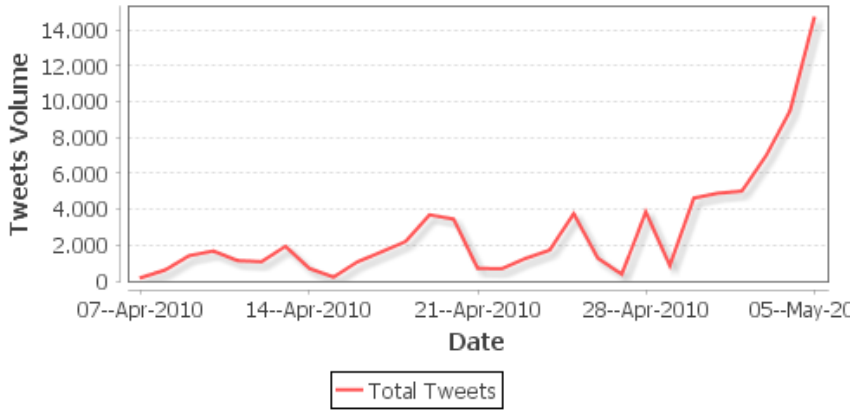


FIGURE 5.17: Tweet volume over 30 days before the elections

Moreover, Figure 5.18 displays the distribution of tweets by different authors across the 30-days period before the elections. The X-axis shows the number of tweets in the log scale, while the Y-axis represents the corresponding frequency of authors in the log scale. The figure shows that the distribution is very close to a Zipfian distribution as a few authors are producing a large number of tweets.

Furthermore, Figure 5.19 shows how the number of tweets per unique author changes over the 30 days for the three most famous political parties. The figure indicates that the ratio remains constant to values between 1 and 1.8 tweets per day during the period of interest.

Computation of Predictor Variables

Figure 5.20 depicts the tweet Volume time series for the Liberal Democrat party. The second time series in the figure is the moving average (MA) of the Volume created by applying the moving average technique over a window of the past k=4 days. A smoothing average over a window of the past k days for variable x is computed as:

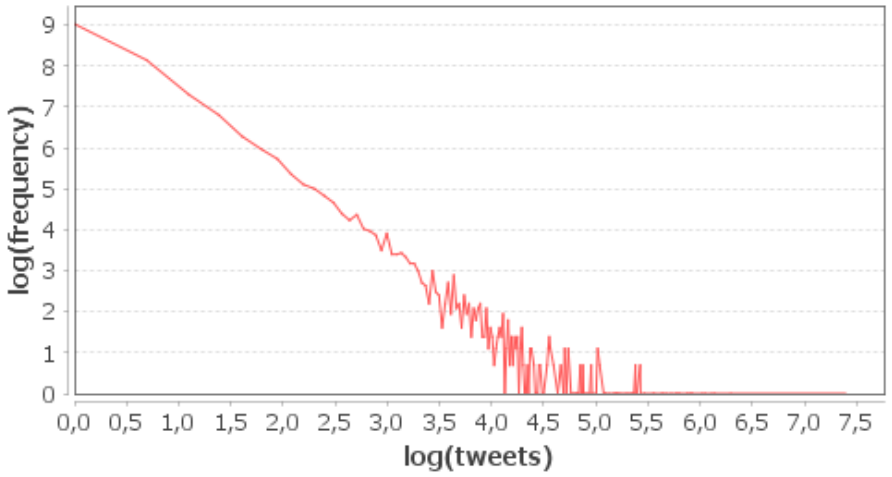


FIGURE 5.18: Log distribution of authors and tweets

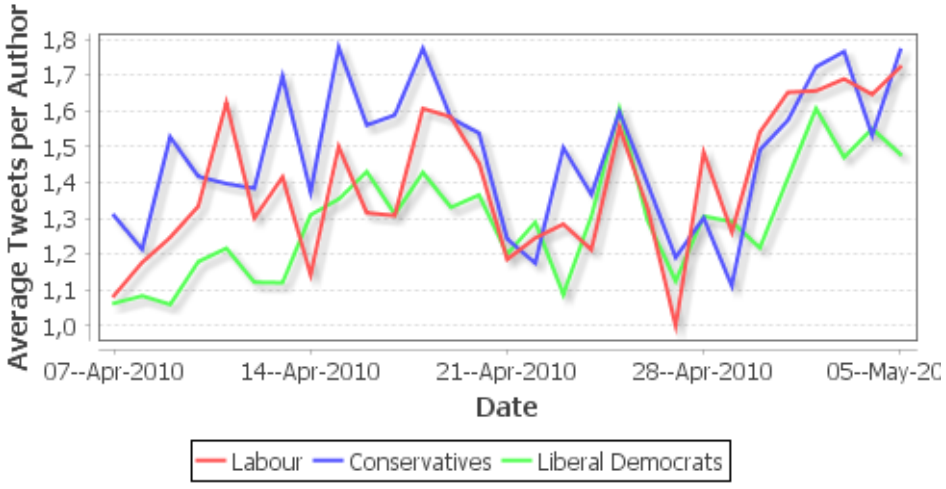


FIGURE 5.19: Average number of tweets per unique author for the three main parties

$$MA_t = 1/k(x_{t-k+1} + x_{t-k+2} + \cdots + x_t) \tag{5.4}$$

Creation of the Predictive Model

Figure 5.8 presents the p-values of the independent variables for the different K windows.

The results of this phase are the regression models created based on YouGov polls as well as the predictor variables. More precisely, the resulting predictive model for the

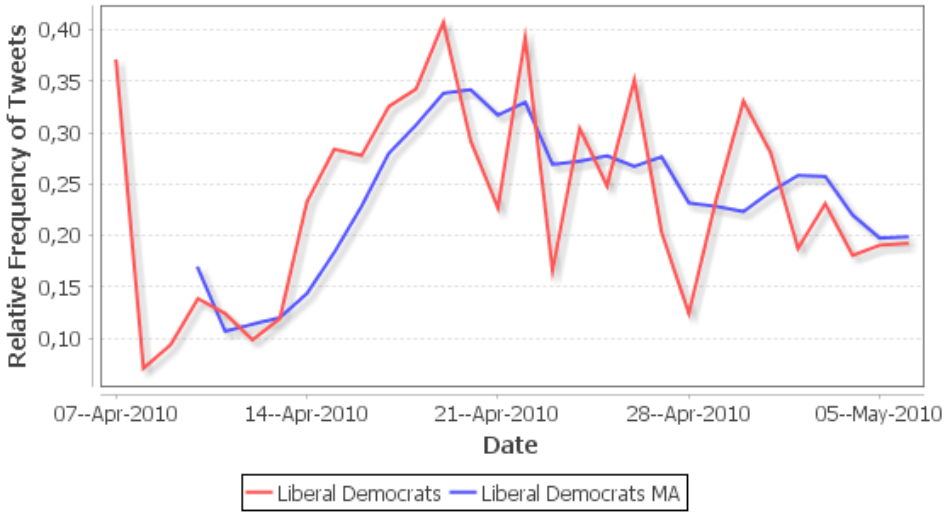


FIGURE 5.20: Volume and Volume Moving Average (k=4) for Liberal Democrats

TABLE 5.8: P-value for the independent variables

	K=4		K=3		K=2	
	RF	*PNR*	*RF*	*PNR*	*RF*	*PNR*
Labours	0.383	0.375	0.458	0.488	0.430	0.583
Conservatives	0.000108 ***	0.357042	1.05e-05 ***	0.0789	5.56e-05 ***	0.0947
Liberal Democrats	1.01e-08 ***	0.38142	6.40e-09 ***	8.737	9.23e-06 ***	0.831

bivariate model that uses RF as predictor variable is described by the following equation:

$$y = 0.54319RF + 0.12812 \tag{5.5}$$

This model is for K=2 and it has been used with training data for the first 20 days, i.e. from 7/4 to 27/4.

Evaluation of the Predictive Performance

We used our regression models to predict the percentage of votes of the Liberal Democrats party. Table 5.9 shows the predicted percentages for the Liberal Democrat party for the 8 days that used as a training set. We can see that the predicted vote percentage using

TABLE 5.9: Prediction results with different models

Date	RF	Prediction	YouGov	Standard Error
28/4	0.163786359	0.2170819	0.310	0.0929181
29/4	0.179767718	0.2257628	0.280	0.0542372
30/4	0.282935266	0.2818023	0.280	0.0018023
1/5	0.305642262	0.2941364	0.280	0.0141364
2/5	0.233985064	0.2552131	0.290	0.0347869
3/5	0.209287988	0.2417979	0.280	0.0382021
4/5	0.205952281	0.239986	0.240	1.4E-05
5/5	0.1857447	0.2290094	0.280	0.0509906
			AVG:	0.03588595

our independent variable has an average prediction error of 0.035 (or 3.5% of votes) when comparing with the forecast that YouGov made.

Interestingly however, all our regression models predicted more accurately the percentage of the Liberal Democrats at the day of the elections. More specifically having in mind that Liberal Democrats gained the 23% of the total UK votes in the elections, our model forecasted 22.9% remarkably better than YouGov's 28% forecast.

5.4.2.4 Discussion

In this section we presented a case for predicting election results using Twitter data. This case falls into the broader category of studies that exploit Social Media data to predict the outcome of real-world phenomena.

The novelties of the presented case in relation to the literature are two fold. The first one regards the approach that is followed in studies aiming at predicting election results. Our study is the first one that is compatible with all suggestions made by a theoretical framework for social media data analysis Kalampokis et al. (2013b) with regards to the data analysis steps followed and the approaches adopted in each step. In particular our study utilizes a dynamic approach to identify keywords for data collection and filtering,

employs both volume and sentiment related variables, computes the sentiment variable through a machine-learning algorithm, uses time-series to create the predictive model, and evaluates the predictive performance using predictive metrics and out-of-sample data. The second novelty of our study is related to the utilization of the Linked Data paradigm to semantically enrich tweets and reuse objective data that is freely available on the Web in order to support the understanding of the tweets. Although this technique has been already proposed in the literature, our study is the first one that adopts it in a case of exploring the predictive power of Social Media.

Two others studies in the literature aim also at predicting the UK election of 2010 through Social Media data Franch (2013); He et al. (2012). Franch Franch (2013) used data from Facebook, Twitter, Google and YouTube and computed numerous variables referring to both volume and sentiment. He also created a linear regression prediction model using poll data from YouGov. As regards the vote percentage of Liberal Democrats, the author found the higher prediction accuracy (average of three bivariate models 23.8%) when used Facebook related variables, while he did not present results for Twitter data. As in our case, the result was close to the real outcome (23%) but it presented a big error in relation to the YouGov poll (28%) that was used to train the model. However, the author does not provide information regarding the collection and filtering stage and the computation of the sentiment variables. On the other hand, He et al. He et al. (2012) collected data from Twitter and applied a lexicon-based approach as well as unsupervised and supervised machine learning algorithms to compute the public sentiment in UK expressed through the tweets for the three major parties. Although, they did not express the variables as time series and they did not evaluate the predictive performance using predictive analytics, they challenged the power of Twitter to predict the results of the UK elections of 2010.

Although we presented in the section only the Linked Data approach in Social Media collection and filtering, we have also performed the same analysis using keywords that we manually defined. The keywords that we have used are presented as hash tags in the table in the Appendix D. These hash tags were the most frequent hash tags referring to one of the three political parties and their leaders in our dataset. We should note that if all the other steps of the analysis remain the same, the prediction accuracy that we receive is very similar to the one that we got with the Linked Data approach. Although this is the case in the specific study, we believe that the approach of linking Twitter data

to other data on the Web has a significant potential particularly in complex cases with numerous interrelated entities. As a result, we will further discuss the limitations that we faced in performing the collection and filtering stage of our analysis using Linked Data.

The first issue that may introduce uncertainty in the approach is the Named Entity Recognition (NER) process. As described in the approach we achieved 85.89 F1 score in the identification and classification of the named entities that were included in the collected tweets. Although this score is quite high we should note that we used a supervised classifier, which requires the creation of a manually annotated dataset of entities and tweets. If it is to widely adopt this approach unsupervised classifiers have to be used and evaluated as well. Moreover, the identification of representations in DBpedia of the named entities that were recognized in the tweets may also introduce uncertainty in our approach. In this process we used DBpedia lookup, although other tools can be also used, e.g Silk Framework (REF), which is a very popular tool for discovering relationships between data items within different Linked Data sources. It is indicative that in our case, out of 6392 distinct named entities identified from the NER process only 3712 were matched with one or more DBpedia resources (i.e. 58% of the entities). We should also note that we have performed the same process using Silk but the results were less accurate. Moreover, in the case where more than one resource was returned by DBpedia lookup, we introduced an extra step that involves manual effort. In particular, we created a case-specific bag of words using the representations in DBpedia of the eight most popular UK political parties and we computed the cosine similarity between the identified resources and this case-specific bag of words. Although the domain-specific bag of words can be also created in other problem areas or cases as well, it may introduce uncertainty and produce poor results. For example, this process was able to identify relationships between "Cameron" entity and dbpedia:David_Cameron, "Cleggmania" entity and dbpedia:Nick_Clegg, "Lib_Dems" entity and dbpedia:Liberal_Democrats, and "UK" entity and dbpedia:United_Kingdom. We should however note that there where cases where entities were connected to a relative but not identical DBpedia entity. For example, our mechanism connected Tony_Blair entity with dbpedia:Premiership_of_Tony_Blair entity and not dbpedia:Tony_Blair because these two DBpedia resources produced almost similar cosine similarity with the entity "Tony-Blair", with dbpedia:Premiership_of_Tony_Blair presenting better results.

Finally, there were cases were the selected lookup result was wrong. We hence used a cosine similarity threshold in order to avoid establishing those irrelevant connections.

We should also note that our case presents another limitation that is related to the volume of the collected tweets. We collected 84,375 tweets using four relevant to the election hash tags, while for example He et al He et al. (2012) collected 919,662 tweets using more hash tags.

In general, a predictive model enables accurate prediction of phenomenon outcomes based on a new set of observations. In election related cases, however, it is difficult to evaluate a model with a new set of election results because one has to wait four or five years until the next election to take place. In this section, we assume that YouGov opinion polling represents an alternative expression of political preference, similar to voting. We hence created our model using data from YouGov and evaluate the model based on the actual election results. We should, however, note, that the YouGov poll of the last day before the elections differ from the actual voting results with Liberal Democrats presenting a difference of 3.4% in the voting percentage. This fact should be carefully analyzed when interpreting the accuracy of such a predictive model.

Chapter 6

Data Access Control on the Web

6.1 Introduction

The objective of this chapter is to study how access constraints affect the reuse of data on the Web and to present a solution based on Linked Data. Towards this end, we present the *Linked Medical Data Access Control (LiMDAC)* framework that capitalizes on Linked Data technologies to enable controlling access to *medical data* across distributed sources with diverse access constraints. The framework consists of (a) three Linked Data models, namely the LiMDAC metadata model for describing aggregated medical data, the LiMDAC user profile model for describing medical data consumers, and the LiMDAC access policy model, and (b) an architecture that exploits and orchestrates the three models to enable controlling access to medical data. From a technological perspective, the framework is validated using a proof-of-concept platform that is developed for that purpose.

Ethical and legal challenges related to medical data mainly derive from (a) strict regulations that protect personal data and prevent patient re-identification by any means [20], (b) agreements that are specified in consent forms, e.g., patients approve sharing their data only in certain clinical studies (Ludman et al., 2010), and (c) policies of stakeholders owing the data e.g., pharmaceutical companies do not contribute to a clinical research led by competitors, or physicians exclude data derived from studies in progress

(Anderson and Edwards, 2010; Dhopeshwarkar et al., 2012; Huang et al., 2009). In general, ethical and legal challenges impose access constraints that can be categorized as follows (Barrows and Clayton, 1996; Fernández-Alemán et al., 2013):

- Secrecy: Ensures the privacy of patients and the confidentiality of medical data preventing unauthorized disclosures of information.
- Integrity: Ensures the integrity of medical data and prevents the unauthorized or improper modifications of data.
- Availability: Ensures the availability of medical data only to authorized persons and prevents the unauthorized or unintended withholding of data.

In order to overcome secrecy and integrity constraints, the approach of aggregating data has been proposed. Aggregated data includes only counts of patients having specific characteristics instead of raw record-level information. Aggregated data is usually structured in the form of multi-dimensional data cubes. In this way, non-identification and anonymization are ensured while the original data remain safe from any modifications. Despite that, there is still a need for controlling access to aggregated data, e.g., due to data provider's policies, and thus availability constraints call for appropriate solutions (Caine and Hanania, 2013).

6.2 Related work

6.2.1 Access control on medical data

Two different approaches to discover and access medical data from multiple data sources have been proposed: centralized and distributed (Weiner and Embi, 2009). The centralized approach enables access to medical data that have been transferred in centralized repositories. Alternatively, distributed medical data networks enable discovery of medical data in their original, disparate locations.

Several frameworks consisting of processes, data models and software tools have been recently developed in order to enable medical data sharing and reuse (Malin et al., 2010). For example, the Informatics for Integrating Biology and the Bedside (i2b2) (Kohane

et al., 2012; Murphy et al., 2010) open source platform and software implementation allow clinical researchers to search across multiple i2b2 sites, find sets of interesting patients, and reuse medical data while preserving patient privacy and ensuring data integrity. While i2b2 created an analytic platform for a single clinical data repository, the Shared Pathology Information Network (SPIN) (Drake et al., 2007) has tackled the problem of cross-institution data sharing across a peer-to-peer network, in which each participating institution maintains autonomy and control of its own data. The Shared Health Research Information Network's (SHRINE) (Weber et al., 2009) was built upon i2b2 and SPIN to enable investigators to search the electronic health records of patients across multiple independent sites. In the same context, the Cross-Institutional Clinical Translational Research (CICTR) (Anderson et al., 2012) framework also extended i2b2 in order to enable federated queries across i2b2 sites. In particular, implementations of these two frameworks allow querying distributed hospitals and display aggregate counts of the number of matching patients. The Biomedical Informatics Research Network (BIRN) (Keator et al., 2009) aggregates imaging, clinical and behavioural data from multiple independent sites using a mediator that re-submits user queries to the relevant sites and aggregates results. The Service-Oriented Interoperability Framework (SIF) (Slaymaker et al., 2008) targets heterogeneous data sources and employs Web Services standards (e.g., SOAP) to query JDBC databases. In this case users should be aware of all data models in order to form appropriate queries. The Federated Utah Research & Translational Health e-Repository (FURTHeR) (Livne et al., 2011) is also based on Service Oriented Architecture and employs Web Services standards in order to perform federated queries across distributed data sources. Finally, the integrating Data for Analysis, anonymization, and SHaring (iDASH) (Ohno-Machado et al., 2011) framework covers many aspects of medical data reuse including annotation, compression, anonymization, information extraction from text, sharing in a privacy-preserving manner, and integration.

6.2.2 Access control on Linked Data

At the same time, several research efforts have been made so far to control access in data published as Linked Data (Kirrane et al., 2013). Initially, access policies were defined for the entire RDF file stored on a web server. Thereafter, it was attempted to apply access policies on parts of the RDF graph (Flouris et al., 2010; Jain and Farkas, 2006; Kagal

et al., 2003). In order to achieve this, the proposed access control frameworks define parts of the RDF graph on which access can be allowed (or denied). These parts are identified by specifying RDF patterns. Whereas the above approaches have primarily focused on RDF patterns, Costabello et al. (Costabello et al., 2012) and Sacco et al. (Sacco et al., 2011) propose access control ontologies over Linked Data. They both employ the SPARQL ASK to determine whether the requester is allowed, or not to access the requested resource. In general, the ASK query form can be used to test whether a query pattern has a solution and returns whether the solution exists. An important issue that may arise is the increase of the overhead produced by evaluating policies in every RDF triple (Abel et al., 2007).

6.3 Access constraints in clinical research

In order to elaborate on availability constraints, a patient-oriented research and an epidemiological study scenario are described below.

In patient-oriented research, clinical researchers search for subjects that meet certain eligibility criteria related to a clinical study. Initially they identify possible data providers and ask them whether data of relevant subjects is included in their patients' database. Data providers having such data and wishing to participate to the specific clinical study have to perform some intensive tasks. First, they check whether the identified subjects can be included according to the study's eligibility criteria. Then, they match the eligible subjects with the patients' consent forms to identify if they can be enrolled to the specific trial. Finally, they confirm that access to the patient data is permitted without violating any access constraints, e.g., when subjects have been recruited for a different trial. If the number of eligible patients is not sufficient, clinical researchers seek for additional subjects from other sources to meet the recruitment target.

In epidemiological studies, clinical researchers perform statistical analyses of medical data in order to conduct secondary clinical research and thus, identify risk factors influencing the occurrence of a pathological process. In order to have accurate and statistically significant results, they need a large number of medical data. To this end, they identify possible data providers and ask for relevant data they can access without violating any access constraints. Data providers wishing to contribute to the specific

clinical research have to perform some intensive tasks. First, they modify their data in order to ensure that their data will be transferred in an anonymized and non-identifiable form. To achieve this, they delete all references to the subject and create aggregated data. Data providers confirm that access to the data is permitted without violating any policies and provide the data e.g., data that is used for studies in progress.

In addition to the scenarios, we interviewed stakeholders working in organizations that participate in the EU funded FP7 Linked2Safety project (Antoniades et al., 2012). In particular, we interviewed five clinical researchers, one data manager and three clinical study managers coming from three healthcare organizations maintaining and using medical data for clinical research, namely the Institute of Neurology and Genetics in Cyprus, the Lausanne University Hospital, and ZEINCRO Hellas S.A., a private Contract Research Organisation in Greece. This exercise resulted in a list of user requirements that are related to availability constraints. This list is presented in Table 6.1.

These scenarios and requirements enable us to identify that two abstract roles related to medical data management are important in clinical research. The data provider creates and keeps medical data regarding patient-specific information in order to organize patients' treatment, or conduct a clinical research. The data consumer discovers subjects meeting certain eligibility criteria for a patient-oriented research, or medical data to perform an epidemiological study.

In addition, these scenarios and requirements enable us to come up with an abstract process that delineates clinical research and consists of the following steps:

1. Data provider modifies and aggregates data in order to ensure anonymity and non-identification.
2. Data consumer searches for data providers
3. Data consumer asks data provider for certain data.
4. Data provider checks whether the requested data is available.
5. Data provider checks whether data consumer is allowed to access the data according to some access constraint policies.
6. Data consumer receives the data.

TABLE 6.1: Requirements related to Availability Constraints

No	*Requirements*
	Requirements regarding the medical data
R1	The content of medical data e.g., data about cancer. The content derives from the variables that have been recorded for each subject in the clinical study.
R2	The data provider of the medical data e.g., a hospital, a pharmaceutical company or a health agency.
R3	The clinical study from which medical data has been derived.
R3a	The title of the clinical study.
R3b	The research topic of the clinical study e.g., children's obesity.
R3c	The purpose of the clinical study e.g., for an epidemiological study.
R3d	The principal investigator of the clinical study.
R3e	The health institution where the clinical study has been conducted e.g., a clinical site, a hospital, a private medical centre, a clinical laboratory.
R3f	The contributors of the clinical study if additional subjects were needed.
R3g	The sponsor of the clinical study e.g., a pharmaceutical company or a public health authority.
R3h	The period of time that the clinical study has been run.
R3i	The location where the clinical study has been conducted. This usually refers to the location of the healthcare institute performed the clinical study.
	Requirements regarding the data consumer
R13	The name of the data consumer.
R14	The location/origin of the data consumer e.g., a country.
R15	The research interest of the data consumer e.g., oncology.
R16	The organization where the data consumer is working.
R16a	The name of the organization.
R16b	The type of the organization e.g., pharmaceutical company.
R16c	The location of the organization.
R16d	The occupation/position in the organization that the data consumer holds e.g., biologist, epidemiologist, or endocrinologist.
R17	The activity that the data consumer needs to perform. It can also be denoted as the purpose of the data consumer.
R17a	The name of the activity/purpose.
R17b	The type of the activity e.g., clinical trial, epidemiological study, publication.
R17c	The topic of the activity e.g., health habits, breast cancer.
R17d	The role of the data provider e.g., clinical researcher.

6.4 The Linked Medical Data Access Control Framework

The proposed Linked Medical Data Access Control (LiMDAC) framework consists of the following:

- Three Linked Data models, namely the LiMDAC metadata model, the LiMDAC user profile model and the LiMDAC access policy model.
- An architecture that exploits and orchestrates the three models to enable controlling access to medical data.

The framework aims at supporting the abstract process of data management in clinical research presented in Section 6.3. In particular, the steps of the process supported by the LiMDAC framework, along with a mapping to those steps presented in Section 6.3, are depicted in Table 6.2. In this table we assume that a platform has been implemented based on the LiMDAC architecture.

The rest of this section is structured according to the main parts of the process and framework. In particular, Section 6.4.1 presents how aggregated medical data should be developed and published as Linked Data. This corresponds to the setup phase of Table 6.2 and is an essential prerequisite for the LiMDAC framework. We note that this is presented here only for clarity. The processes, technologies and tools for performing these tasks are outside the scope of this chapter, as they are well documented in the relevant literature, e.g., (Perakis et al., 2013). Subsection 6.4.2 describes the LiMDAC metadata model, while Subsection 6.4.3 presents the LiMDAC user profile model. Subsection 6.4.4 elaborates on the LiMDAC access policy model and Section 6.5 describes an architecture that exploits these three models.

In order to enhance clarity, we present a research study about childhood obesity [66]. According to the study, six paediatric academic health sites from different regions of the United States have participated in a clinical research related to children's obesity. The dataset maintained by each site involves records about children between 2 and 17 years old. For each child, the health sites store information about the age and the Body Mass Index (BMI). Moreover, the health sites have detected groups of conditions that most commonly co-occur with obesity including hypertension, hyperlipidaemia as well as other rare disorders such as acute leukaemia, multiple sclerosis, and chromosomal

TABLE 6.2: Abstract clinical research process supported in the LiMDAC framework

Existing Abstract Process	*Abstract Process in the LiMDAC framework*
Setup phase	
1. Data provider modifies and aggregates data in order to ensure anonymity and non-identification.	1. (This is not supported by the current version of the LiMDAC framework). • Data providers publish aggregated data as Linked Data and they use the LiMDAC metadata model to describe them. • Data providers define access constraints using the LiMDAC access policy model. • Data consumers create profiles based on the LiMDAC user profile model.
Access phase	
2. Data consumer searches for data providers.	2. Data consumer searches for data providers' SPARQL endpoints through the LiMDAC platform.
3. Data consumer asks data provider for certain data.	3. Data consumer search for suitable data based on the LiMDAC metadata model.
4. Data provider checks whether the requested data is available.	4. The LiMDAC platform checks whether the providers' datasets include suitable data.
5. Data provider checks whether data consumer is allowed to access the data according to some access constraint policies.	5. The LiMDAC platform checks data consumer's profile against the available access policies.
6. Data consumer receives the data.	6. Data consumer receives suitable data through the LiMDAC platform.

TABLE 6.3: Children obesity data cube coming from a data provider

Disease	Hypothyroidism				Diabetes			
BMI	15	16	17	...	15	16	17	...
Age								
2	22	22	23		12	23	11	
3	23	22	23		20	23	12	
4	22	22	24		30	24	22	
...								

anomalies. In this chapter, we use the background scenario of this study in order to present our results. We should, however, underline that we have used dummy and not real data from the study.

6.4.1 Linked medical data cubes

Based on the obesity example, Table 6.3 depicts part of a truncated dummy data cube provided by one of the sites.

Figure 6.1 presents the RDF graph produced from the data shown in Table 2. The graph is modelled based on the RDF data cube vocabulary. In particular, it describes the data structure definition, along with two observations. The SKOS concept collection is used to indicate a set of disease concepts. Figure 6.1 also presents the links that have been established between concepts and external vocabularies. It is apparent that *ex:disease* is linked to the concept *EFO:disease* from the EFO ontology that has the same meaning. The concept *EFO:bodymassindex* can be reused for the BMI dimension. Moreover, the measurement of frequencies and the age dimension can be expressed using the Statistical Data and Metadata eXchange standard (SDMX) which is used to publish statistical data on the web.

We repeat here that medical data is transformed in this format during the setup phase shown in Table 6.2, is a prerequisite for employing the LiMDAC framework, and is not further elaborated in this chapter.

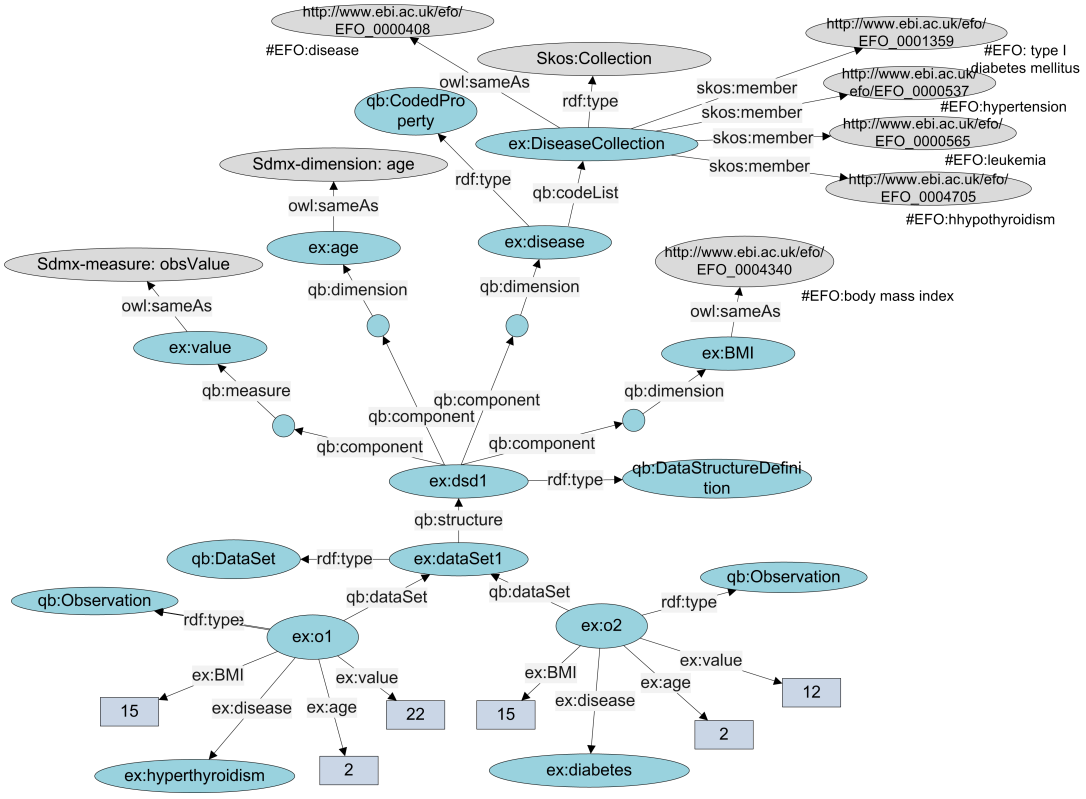


FIGURE 6.1: RDF graph of Children's Obesity Data Cube

6.4.2 The LiMDAC metadata model

The LiDMAC metadata model facilitates the improved description of linked data cubes. Metadata enable data providers to express richer access constraints and data consumers to perform more expressive search queries.

The RDF data cube vocabulary includes the *qb:DataStructureDefinition* concept which defines metadata related to cube structure. These metadata include the dimensions of the data cube, along with the values that are measured. However, additional metadata are needed to provide information about the clinical study (e.g., title, purpose, duration, location, subject, responsible personnel for conducting the clinical study etc.) and the aggregation process that has been followed. These metadata were extracted from the requirements described in Table 6.1.

Table 6.4 presents the mapping between the requirements and the concepts extracted for describing medical data cubes. Based on these concepts, a conceptual model has been created, which is depicted in Figure 6.2.

TABLE 6.4: Concepts of the LiMDAC metadata model as elicited from user requirements

Concepts	*Data*	*Study*	*Agent*	*Role*	*Variable*	*PeriodOfTime*	*Location*
R1	✓				✓		
R2	✓		✓	✓			
R3	✓	✓					
R3a		✓					
R3b		✓					
R3c		✓					
R3d		✓	✓	✓			
R3e		✓	✓	✓			
R3f		✓	✓	✓			
R3g		✓	✓	✓			
R3h		✓				✓	
R3i		✓	✓				✓

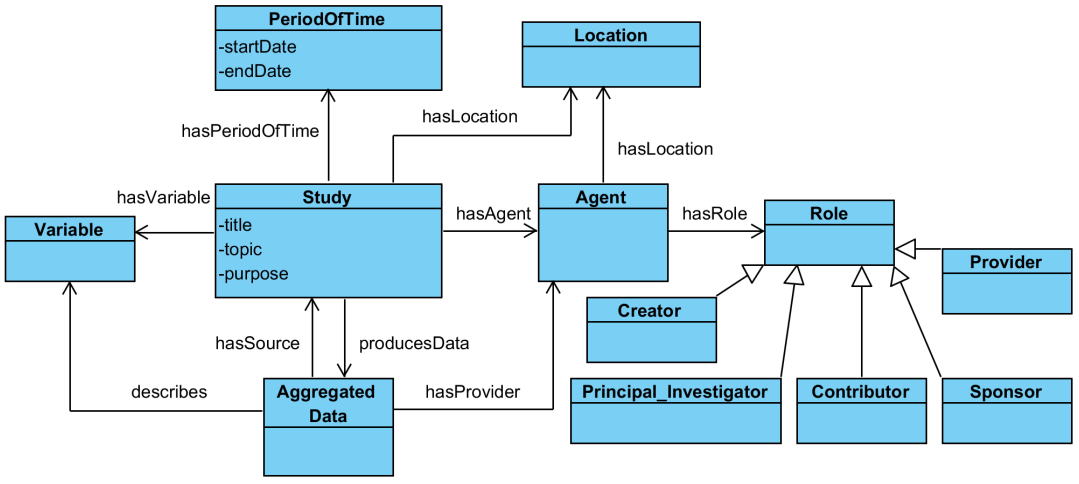


FIGURE 6.2: The LiMDAC metadata conceptual model

Figure 6.3 shows the the LiMDAC metadata model in the form of a linked data vocabulary. Since the model follows the linked-data principles, it reuses concepts from existing vocabularies instead of defining new ones. In particular, the following popular linked data vocabularies are exploited:

- The DDI Discovery Vocabulary (Bosch et al., 2013) that describes research and survey datasets on the Web.
- The DCMI Metadata Terms vocabulary [1] that is a specification of all metadata terms used to describe a resource.
- The FOAF vocabulary [2] that describes people and their relationships.
- The SKOS vocabulary [3] that is used to define classifications.

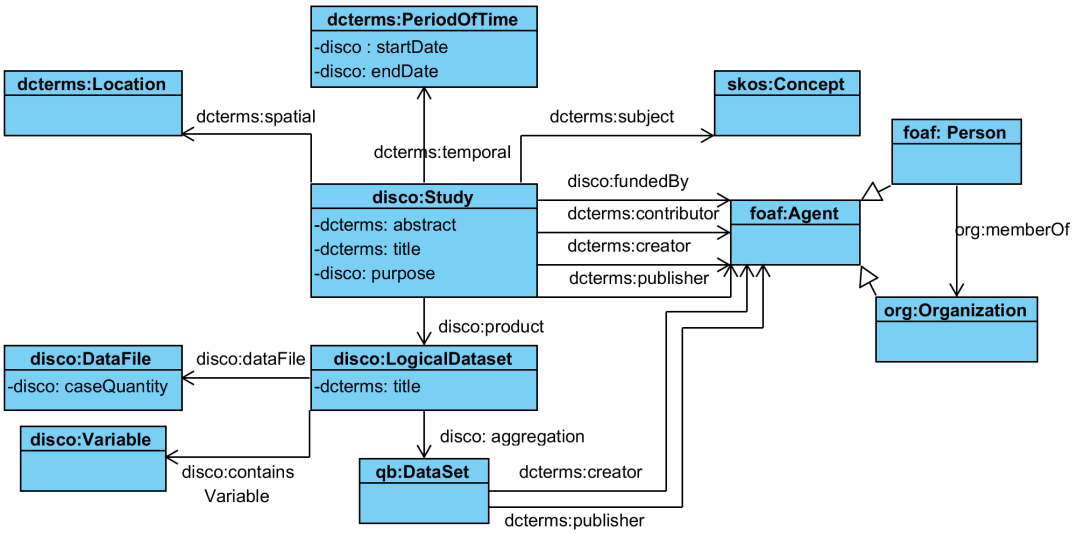


FIGURE 6.3: The LiMDAC metadata linked data model

In order to describe the proposed metadata model, we use the Study class (*disco:Study*) as an entry point. A Study represents the process by which a dataset was collected. It may contain a set of literal properties that provide information about the title (*dcterms:title*), the purpose (*disco:purpose*), and other high-level information. In addition, a study includes object properties such as the sponsor of the study (*disco:fundedBy*), and other affiliations such as creators, contributors and publishers of Studies (*dcterms:creator*, *dcterms:publisher*, *dcterms:contributor*).

[1] http://dublincore.org/documents/dcmi-terms/
[2] http://xmlns.com/foaf/spec/
[3] http://http://www.w3.org/TR/skos-reference/

They all are *foaf:Agents*, which can be either *foaf:Persons* or *org:Organizations* whose members are *foaf:Persons*. Moreover, we can use the subject, the temporal and the spatial properties (*disco: subject*, *dcterms:temporal*, *dcterms:spatial*) to describe the respective coverage of studies. For time periods, start (*disco:startDate*) and end dates (*disco:endDate*) can be also attached.

The outcome of a study is a raw record-level data set (*disco:LogicalDataSet*) that may point to multiple variables (*disco:Variable*). The LogicalDataset contains a property disco:aggregation that indicates that a data cube (*qb:Dataset*) was derived by tabulating a LogicalDataset. Furthermore, the Study/LogicalDataset has a DataFile (*disco:DataFile*) that is the distributed file holding that data.

In the example of childhood obesity, the Children's Hospital of Colorado is the creator of the study that is entitled "Multi-institutional study to access childhood obesity" and its purpose is to associate several health conditions with childhood obesity. Michael G. Kahn is the principal investigator of the study as well as responsible person for creating the data cubes transforming the produced medical data in an aggregated form and publishing them as linked data. The study was funded by the Agency for Healthcare Research and Quality (AHRQ) and run from January 2007 to December 2008.

Taking into account all this information, the Children's Hospital of Colorado enriches the linked data cube derived from this study with the following LiMDAC metadata as shown in Figure 6.4.

6.4.3 The LiMDAC User Profile Model

The LiMDAC user profile model is used to describe data consumers. This model is exploited by data providers to define their access constraints and by data consumers to describe their user profiles. The LiMDAC user profile model should be in alignment with the requirements expressed by clinical research stakeholders (Table 6.1). Table 6.5 presents the mapping between users requirements and the identified concepts that are used in the model (Figure 6.5).

Figure 6.6 shows the proposed LiMDAC user profile model in terms of a linked data vocabulary. Again, the model capitalizes on popular linked data vocabularies by reusing

```
:childhoodObesityStudy a disco:Study;                    #metadata about the study
    dcterms:title          "Multi-institutional study to access childhood obesity";
    disco:purpose          "Identify the risk factor of childhood obesity";
    dcterms:temporal       [
      disco:startDate      "2007-01-01";
      disco:endDate        "2008-12-31";]
    dcterms:spatial        :Colorado;
    dcterms:creator        :ChildrenHospitalColorado;
    disco:fundedBy         :AgencyforHealthcareResearchandQuality;
    disco:product          :childhoodObesityLogicalDataSet.
:childhoodObesityLogicalDataSet  a disco:LogicalDataSet; #metadata about the logical dataset
    disco:dataFile         :obesityLogicalDataFile;
    disco:aggregation      :childhoodObesityDataCube1, :childhoodObesityDataCube2, ..
:obesityLogicalDataFile a disco: DataFile;               #metadata about the data file
    disco:caseQuantify     "1000";
:childhoodObesityDataCube1 a qb:DataSet;                 #metadata about the data cubes
    dcterms:creator        :MichaelGKahn;
    dcterms:publisher      :MichaelGKahn;
    qb:structure           :childhoodObesityDataStructureDefinition1 #structure's dimensions:disease,
                                                                      BMI and age
:childhoodObesityDataCube2 a qb:DataSet;
    dcterms:creator        :MichaelGKahn;
    dcterms:publisher      :MichaelGKahn;
    qb:structure           :childhoodObesityDataStructureDefinition2 #structure's dimensions:BMI, age
                                                                      and physical activity
[...]
```

FIGURE 6.4: Metadata of Children's Obesity Data Cube

TABLE 6.5: Concepts of the LiMDAC user profile model as elicited from user requirements

Concepts	*Person*	*Organization*	*Position*	*Location*	*Activity*	*Role*
R13	✓					
R14	✓			✓		
R15	✓					
R16	✓	✓				
R16a		✓				
R16b		✓				
R16c		✓		✓		
R16d		✓	✓			
R17	✓				✓	
R17a					✓	
R17b					✓	
R17c					✓	
R17d					✓	✓

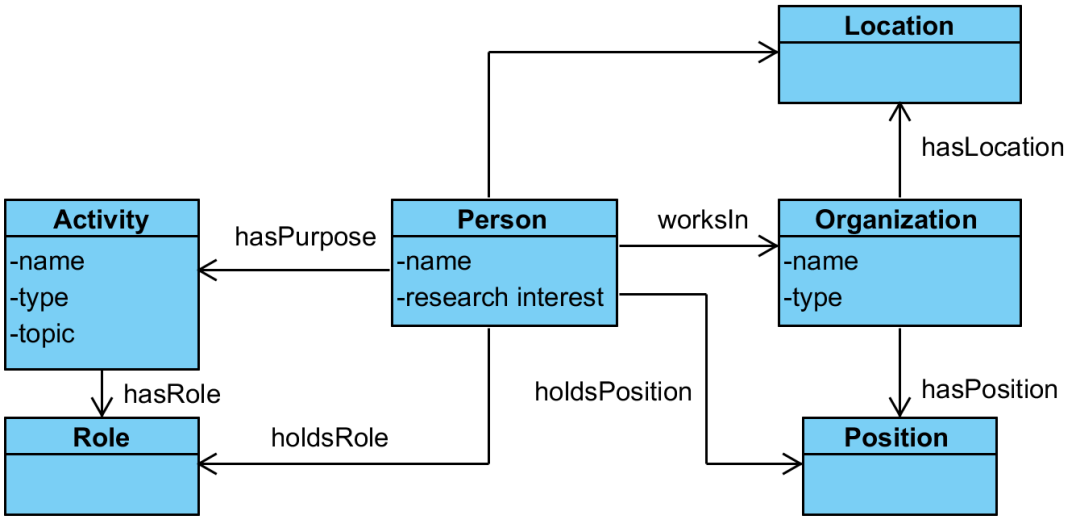


FIGURE 6.5: The LiMDAC user profile conceptual model

existing concepts instead of defining new ones. In particular, it capitalizes on the following linked data vocabularies:

- The FOAF vocabulary that defines the agent.
- The Organization ontology [4] that is used to describe organizational structures.
- The PROV ontology [5] that is used to model provenance information.
- Finally, the SKOS vocabulary that is used to define classifications.

We use the Organization vocabulary to define the Organization (*org:Organization*) that a data consumer works for (*org:memberOf*). The class org:Post represents the position that the data consumer holds in the Organization. The *org:Site* denotes the office or other premise at which the Organization is located. In addition, the Site uses the property *org:siteAddress* to indicate the address of the Site. The SKOS vocabulary is used to define the classification of the Organization within some classification scheme (*org:classification*). Furthermore, the PROV concept *prov:Activity* associates an agent (*prov:Agent*) with a action/activity that he plans, or is responsible to conduct on the extracted data.

In the example of childhood obesity, we could consider two clinical researchers, *researcher A* and *researcher B*.

[4] http://www.w3.org/TR/vocab-org/
[5] http://www.w3.org/TR/prov-o/

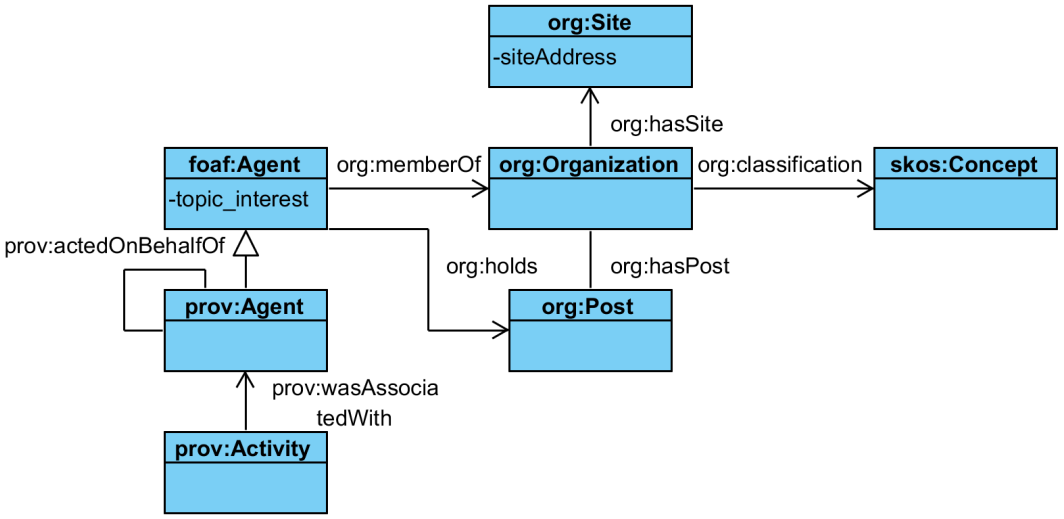


FIGURE 6.6: User metadata model

Researcher A works as a biologist. His current occupation is principal trial investigator in AHRQ while his latest research interests are related to the effects of the obesity in children's life. Researcher A has decided to undertake a clinical research related to the effectiveness of a new treatment to children obesity for children presenting diabetes. To this end, he needs a sufficient number of subjects aged under 18 and diagnosed with diabetes (patient-related research).

Researcher B works as an endocrinologist in the Children's Hospital of Philadelphia. His research interests are related to health habits. Recently, researcher B has started an epidemiological study about health habits and their effects on children's health. In order to receive statistically significant results from the statistical analyses, he needs a sufficient amount of medical data derived from youth patients aged under 18 and containing information about BMI, regular exercise and/or vegetable consumption (epidemiological study).

Taking into account all this information, the user profiles of the two data consumers are presented in Figure 6.7.

6.4.4 The LiMDAC Access Policy Model

Access policies define the data to be protected and to whom access is granted or denied. Thus, each access policy consists of two parts. The first includes the metadata profile

```
:ResearcherA a foaf:Agent, provAgent;                      #metadata about the data consumer A
    foaf:topic_interest     :ChildrenObesity;
    org:memberOf            :AgencyforHealthcareResearchandQuality;
    org:holds               :Biologist.
:PatientOrientedResearchforDiabetes a prov:Activity;       #metadata about the purpose/activity
    prov:wasAssociatedWith  :ResearcherA;
    rdfs:comment            "ResearcherA should perform a patient-oriented research about obesity for
                            children having diagnosed with diabetes".
:AgencyforHealthcareResearchandQuality a org:Organization;  #metadata about the organization A
    foaf:homepage           "http://www.ahrq.gov/";
    org:hasSite             [org:siteAddress "540 Gaither Road, Rockville, MD 20850";]
    org:classification      :FederalAgency.
:ResearcherB a foaf:Agent, prov:Agent;                     #metadata about the data consumer B
    foaf:topic_interest     :HealthHabits;
    org:memberOf            :ChildrenHospitalPhiladelphia;
    org:holds               :Endocrinologist.
:EpidemiologicalStudyforHealthHabits a prov:Activity;      #metadata about the purpose/activity
    prov:wasAssociatedWith   :ResearcherB;
    rdfs:comment             "ResearcherB should perform an epidemiological study about health habits
                             and their effects on children's health".
:ChildrenHospitalPhiladelphia a org:Organization;          #metadata about the organization A
    foaf:homepage           "http://www.chop.edu/";
    org:hasSite             [org:siteAddress "34th Street and Civic Center Boulevard, Philadelphia,
                            PA 19104";]
    org:classification      :Hospital.
:HealthOrganization a skos:Collection                      #A skos collection for organizations
    skos:member :Hospital, :ClinicalSite, :ResearchInstitute, :PharmaceuticalCompany, :FederalAgency...
```

FIGURE 6.7: User profiles

of the medical data that will be protected and the second the profile of data consumers that are allowed (or not) to have access to the data.

In the LiMDAC framework we adopt a simplified access policy approach that enables us to assign access policies dynamically on linked data cubes sharing common characteristics. In particular, the LiMDAC access policy model specifies a) an RDF pattern based on the LiMDAC metadata model to limit the application of policies only to data cubes annotated with those metadata and b) a user pattern based on the LiMDAC user profile model to give the access permission only to users described with those attributes.

Figure 6.8 depicts the LiMDAC access policy linked data model that consists of the following concepts:

- The Dataset: It defines the dataset, where the access policy is applied. The dataset is usually a store containing all linked data cubes of a provider.
- The Data Cube Space: It describes the data cubes, in which the access policy applies. This is achieved through an RDF pattern based on the LiMDAC metadata model that should be satisfied by the metadata of a data cube. If the metadata contain this pattern then the access policy is applied to the data cube.

- The Access Space: It defines the data consumers for which the access policy applies. This is achieved through an RDF pattern based on the LiMDAC user profile model that specifies a user profile.

- The Access Control Privilege: It defines both types of permissions (i.e., grantAccess/denyAccess) and permitted operations (read/write/update). We define it as a subtype of the acl:Access.

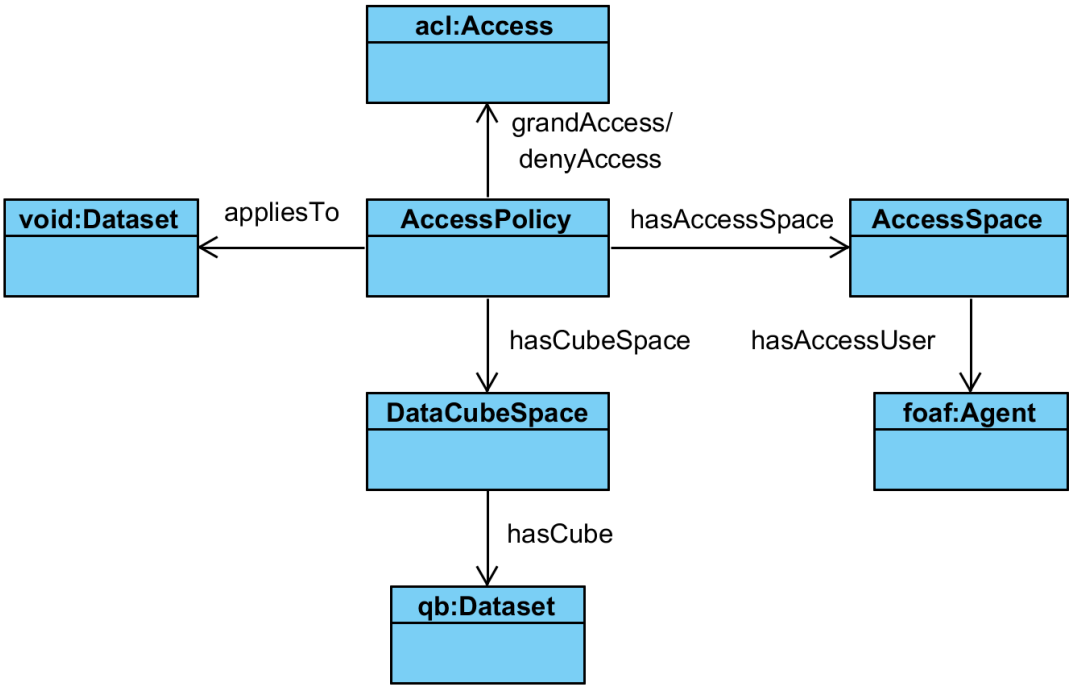


FIGURE 6.8: Access policy model

In the example of childhood obesity, each paediatric hospital serves as an individual data provider maintaining in a local repository a large number of linked data cubes coming from various clinical researches. Based on the proposed access policy model, each data provider creates access policies to make available the linked data cubes only to authorized data consumers and under specific conditions. Consider the case of the Children's Hospital of Colorado. Using the proposed access policy model, it creates the following access policies:

- The data cubes derived from the childhood obesity study and sponsored by AHRQ can be accessed by any user working at AHRQ.

- The data cubes having in their structure the BMI dimension are authorized for access to data consumers working at the Children's Hospital of Philadelphia and being endocrinologists.

Figure 6.9 presents the RDF representation of the second access policy.

```
:AccessPolicy2 a :AccessPolicy;
    rdfs:label          "This access policy enables data consumers working at Children's Hospital of
                         Philadelphia and being endocrinologists to have 'read' access to linked
                         data cubes having in their structure the BMI dimension";
    :grandAccess        acl:read;
    :hasCubeSpace       [
                        :cubeQuery "ASK {
                        ?cube <http://purl.org/dc/elements/1.1/creator> : ChildrenHospitalColorado.
                        ?cube <http://purl.org/linked-data/cube#structure> ?struct.
                          ?struct <http://purl.org/linked-data/cube#component> ?comp.
                            ?comp <http://purl.org/linked-data/cube#dimension> :BMI.}"^^xsd:string.
     :hasAccessSpace    [
                        :accessQuery  "ASK {
                        ?x <http://www.w3.org/TR/vocab-org/#org:holds> :Endrocrinologist.
                        ?x <http://www.w3.org/TR/vocab-org/#org:memberOf>
                            :ChildrenHospitalPhiladelphia.}"^^xsd:string.
```

FIGURE 6.9: Example of access policy

6.5 The LiMDAC Architecture

Figure 6.10 illustrates an architecture that exploits and orchestrates the three LiMDAC models to enable controlling access to medical data. Apart from the LiMDAC models, the architecture includes an Authorization Mechanism module and an Authorization Interface module.

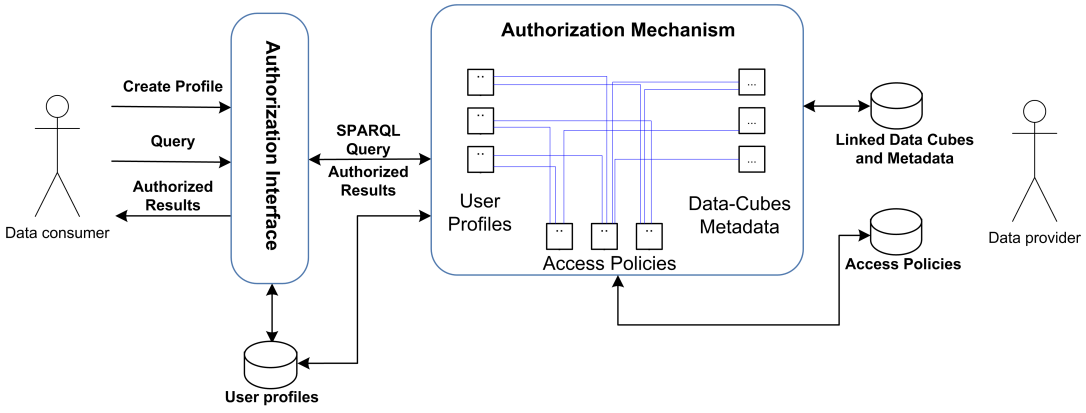


FIGURE 6.10: The LiMDAC solution

The Authorization Interface lies between data consumers and providers. It enables data consumers to i) create user profiles based on the LiMDAC user profile model and ii)

search for distributed medical data. The data consumer defines the purpose of accessing medical data being either patient-oriented research or epidemiological study. The search criteria are based on the dimensions of the data cubes that are stored in the distributed RDF data stores. Next to each search criteria there is a field for selecting its value from a drop-down list of available codes.

Moreover, the Authorization Interface translates data consumers' queries into SPARQL queries and passes them on to the Authorization Mechanism. Initially the Authorization Mechanism retrieves the access policies from distributed data providers. Thereafter, the Authorization Mechanism checks whether the profile of the data consumer matches the user profile defined by each access policy. In case of success, the Authorization Mechanism creates and sends SPARQL queries to distributed data providers. The queries search for data that match both the data consumer's query and the satisfied access policies. In case of success, the resulted datasets are returned to the data consumer via the Authorisation Interface.

The result of this process is either i) (for patient-oriented research purposes) the number of patients meeting specific criteria, along with the name of the data provider publishing these data, or ii) (for epidemiological studies purposes) the respective linked data cubes.

6.6 Proof-of-Concept Implementation and Evaluation of the Platform

Based on the framework a proof-of-concept platform is developed and its functionality and performance are evaluated based on two usage scenarios. The platform implements the Authorization Mechanism and the Authorization Interface and exploits the LiMDAC models. For its development we used the Jena Framework.

Following on the childhood obesity example, we consider that there are four distributed data providers, each corresponding to a different hospital. Each data provider stores linked data cubes produced in the course of different clinical studies (including the childhood obesity example). Data providers also store metadata and access policies for the cubes based on the LiMDAC metadata and access policy models respectively. To emulate this setting, we have used four distributed RDF data stores (implemented using

a Fuseki SPARQL server (part of the Jena Framework). A total of 120.000 linked data cubes are stored in each site. Each cube is structured based on 3 out of 10 dimensions that have been selected for each clinical study. We use around 40 dimensions in total with each receiving values from a pre-defined coded list. Moreover, there exist around 1.5 million triples expressing data cubes' metadata that will be searched by access policies.

In order to evaluate the platform we investigate two usage scenarios. In the example of childhood obesity, researcher A (as defined in Section 4.3) needs medical data for patient-oriented research (purpose) and selects the age dimension with value smaller than 18 and the diabetes dimension. On the other hand, researcher B needs medical data for epidemiological purposes and selects the BMI, Regular_Exercise, Vegetable_consumption and age dimensions. In the value field of the age dimension he indicates smaller than 18 because so that to investigate effects of health habits in young people.

In the first scenario, the system returns the number of subjects meeting the specified criteria. Specifically, Researcher A is now provided access to data derived from the childhood obesity study, which his organisation has funded.

In the second scenario, the system returns data cubes meeting all specified criteria. We assume that this data comes from the Children's Hospital of Philadelphia and the Children's Hospital of Colorado. Researcher B works at the Children's Hospital of Philadelphia, thus he is provided access to data coming from his organization. Moreover, he is provided access to data coming from the Children's Hospital of Colorado, since the Children's Hospital of Colorado has defined a special access policy so as endocrinologists coming from the Children's Hospital of Philadelphia are permitted access to data cubes related to BMI.

In order to evaluate the proof-of-concept platform, we performed functionality, as well as performance and scalability, testing based on two usage scenarios. In the functionality evaluation testing, we validated and verified that all requirements are implemented by the LiMDAC framework and the proof-of-concept platform.

In order to evaluate the performance and scalability of the platform, we have conducted two sets of experiments measuring the response time. At each running, we have executed a complex (3 search criteria) and a simple (1 search criteria) user query. Table 6.6 presents an example of these queries. The main difference between them is the number

TABLE 6.6: Complex and simple queries

Complex Query Example

```
SELECT ?diabetes ?bmi ?hyper
WHERE {
 ?cube qb:structure ?struct
 ?struct qb:component ?comp
 ?comp qb:dimension :Diabetes
 ?comp qb:dimension :BMI
 ?comp qb:dimension :Hypothyroidism
}
```

Simple Query Example

```
SELECT ?cube
WHERE {
 ?cube qb:structure ?struct
 ?struct qb:component ?comp
 ?comp qb:dimension :Diabetes
}
```

of dimensions requested. In our example, diabetes, BMI and hypertension dimensions are requested in the complex query while diabetes is requested in the simple one:

At the first experimental setting (Table 6.7), each provider has 1,000 access policies (we assume that 10% of the access policies are satisfied by the user profile). The measurement is repeated for 1,200, 12,000 and 120,000 cubes per provider. The response time varies from 1 sec to 23 sec for the simple user query while for the complex query the response time are much sorter (vary from 0.1 sec to 0.6 sec). The more search criteria are used the less cubes are matched and thus, the quicker the response is. At the second experimental setting (Table 6.8), the number of cubes is constant, namely 120,000. Here, the variable parameter is the number of access policies per provider. We repeated the measurement for 10, 50, 100, 1,000, 5000 and 10,000 access policies per provider (assuming that 10% of the access policies are satisfied). Again at each case a simple and a complex user query are executed. For the simple query the response times are between 0.1 sec and 23 sec for realistic scenarios until 1000 access policies, while the response time increases significantly as the access policies increase to simulate a web scale. For the complex query the response times are between 0.1 sec and 0.3 sec for realistic scenarios (i.e., until 1000 access policies) while the response time reaches 5 sec and 12 sec for imaginary scenarios that simulate a web scale (5000 and 10000 access policies).

Based on the performance evaluation results, we observe that the authorization mechanism is more sensitive to the number of available access policies than to the number of available cubes. At a real word scenario, it is expected to have a huge number of cubes

TABLE 6.7: First experimental setting with 1,000 access policies per provider

Number of cubes	*Simple query time (ms)*	*Complex query time (ms)*
1,200	832	119
12,000	3,816	262
120,000	23,249	657

TABLE 6.8: Second experimental setting with 120,000 cubes per provider

Number of access policies	*Simple query time (ms)*	*Complex query time (ms)*
10	180	150
50	1,672	297
100	2,117	319
1,000	23,249	657
5,000	562,141	5,438
10,000	1,547,029	12,239

but the number of policies is not expected to be so high. Based on that, we conclude that the proposed approach is expected to work efficiently for web scale data.

6.7 Conclusion

Several frameworks have been recently proposed to enable sharing and reuse of medical data. These frameworks consist of data models, processes, architectures, and software tools to address various challenges of the data value chain, including ethical and legal ones. These challenges impose access constraints that are related to (a) the privacy of patients, (b) the integrity of data, and (c) the availability of data to authorized only persons. Although existing frameworks provide adequate solutions to the first two constraints, they usually follow simple approaches to the latter type of constraints. For example, a common manner to grant access to authorized users is through a static list of IP addresses or particular people.

In this chapter, we presented the Linked Medical Data Access Control (LiMDAC) framework that capitalizes on Linked Data technologies to enable controlling access to linked aggregated medical data across distributed sources with diverse access constraints. Although, the main focus of LiMDAC is access control in a Web-based environment, it caters for all diverse requirements of medical data sharing.

Medical data sharing frameworks usually employ the creation of new data out of the original medical records in order to ensure data integrity. For example, in the i2b2 framework a copy of the medical record is created and thus investigators are free to clean and manipulate it for their own purposes with other i2b2 software. These frameworks also employ techniques that prevent unauthorized disclosure of patients' information or identity. For example, a popular technique concerns returning aggregate numbers of patients that satisfy a query to the record-level medical data. In LiMDAC, data cubes that contain aggregate numbers of patients are created from the actual medical data in order to ensure both the privacy of the patients and the integrity of the data. Although the creation of the data cubes is out of the scope of the current LiMDAC version, the details of creating and querying cubes in a linked data environment are described.

The scope of the existing frameworks ranges from single data repositories to multiple distributed data sources inside an institution or at a cross-institutional setting. A suitable data model is critical in these cases so that medical record data and clinical trial data can fit together and thus diseases, genes, and outcomes can be related to each other. For example, the i2b2 star schema data model is used to instantiate at a project level the raw medical record data. In the case of LiMDAC the RDF data cube vocabulary, which is a W3C standard, is used to model the aggregated data. In addition, the adoption of the Linked Data paradigm enables reuse of existing widely used vocabularies, datasets and code lists and thus maximizes interoperability and alleviates the burden of semantic alignment. This enables users of LiMDAC to perform queries in multiple sites without having to be aware of the underlying schema of the other sites. The majority of recent frameworks is based on a Service Oriented Architecture and employ Web Services standards such as SOAP protocol (e.g., in FURTHeR and i2b2) and paradigms such as the Enterprise Service Bus (e.g., in iDASH framework). LiMDAC is to the best of our knowledge the first framework that adopts the Linked Data paradigm to develop the underlying infrastructure for sharing and reusing medical data. This will enable the easy integration of medical data with other data on the Web (both medical and third

party data e.g., government data). This is expected to enhance the possibility to gain innovative insights in epidemiological studies.

Although several Linked Data access control frameworks have been recently proposed, they all suffer from several shortcomings when applied to RDF data cubes that represent aggregated medical data. Medical data providers assign access constraints to cubes, which are frequently updated, even once per week. So, there is a need to assign access constraints dynamically based on domain specific metadata. However, current Linked Data access control frameworks do not support this need. Moreover, most of these approaches restrict parts of the RDF graph having specific RDF characteristics (e.g., triples containing a particular property), or associate access policies to specific RDF data. An RDF data cube can be considered as a small RDF graph made up of many triples. Thus, access policies should apply on the data cube granularity level and access should be restricted based on cubes' metadata instead of RDF properties. LiMDAC satisfies these requirements by employing the SPARQL ASK form in cubes' metadata. It is also important to mention that the LiMDAC framework has been proposed as a simplified solution to address availability constrains on linked medical data cubes.

From a technical perspective, the results of our initial performance evaluation are promising, as they show only a small increase in query processing time and a linear increase as the number of data cubes and satisfied access policies grows. A significant delay has been noticed for simple queries only when the number of access policies exceeds a specific level (5000 and 10000). However, this is not a realistic scenario according to interviews conducted with clinical stakeholders for extracting requirements about availability constraints. Indeed, they suggested that an average of 5 to 20 access policies will be created for each dataset of cubes.

Chapter 7

Conclusions

This chapter summarises the research contributions of this Thesis and discusses possible directions for future research.

7.1 Summary of the Thesis

Public sector produces, collects, maintains and disseminates a wealth of information. The availability of this information (government data onwards) in easily accessible digital format makes it possible to re-use it and combine it with other digital content to create new added-value services and products. It is widely recognised that such data-based, added value services and products increase government transparency, improve public administration's function, contribute to economic growth and provide social value to citizens. A recent evaluation of the European Directive underpins a number of barriers towards the full exploitation of government data. This situation seems to change in the last years, where a large number of governments worldwide started to massively make data available on the Web. This *Open Government Data* (OGD) movement follows the *Open Data* philosophy suggesting making data freely available to everyone, without limiting restrictions.

The aim of this thesis is to study the recently emerged Open Government Data (OGD) movement. More precisely, our study is focused on exploring whether and how the OGD movement can realise the potential of government data. Towards this end, we study

OGD in a holistic approach by taking into account the viewpoints of both providers and consumers of OGD.

Regarding OGD provision, we studied OGD portals as part of public sector and as such we consider that they inherit deficiencies coming from the decentralised organisational structure of public sector, which comprises multiple administrative levels and functional areas. As a result, we focus on organisational challenges related to OGD provision. From a technological point of view we focus on *Linked Data* as a paradigm that facilitates data integration on the Web. Linked data has been early proposed as the most advanced and promising way for opening up OGD. As a result, we adopted linked data as the technological paradigm that underpins our implementations and guides our technological analyses. Moreover, we emphasised on aggregated statistics (e.g. economic and social indicators) structured as multi-dimensional data cubes because these data constitute a significant part of the available open data provided through OGD portals. It is indicative that the vast majority of the datasets published on the open data portal of the European Commission are of statistical nature.

Regarding OGD exploitation, on this thesis we focused on creating added-value through data integration and analytics. We studied OGD integration both inside and outside public sector. In the fist case, we studied the integration of data from public agencies in different administrative levels or functional areas that, however, refer to same real world problem. In the second case, we considered data from other sources on the Web. In particular, we exploited data from (a) social media data, (b) news media, (c) clinical researches, and (d) DBpedia, i.e. the linked open data version of Wikipedia. We also employed data analytics methods in order to create value from integrated data. In this thesis we focused on analysing data using typical methods that are used for years in business intelligence and analytics. We exploit online analytical processing (OLAP) as well as well established statistical analysis and data mining techniques for association analysis, classification and regression analysis, and predictive modelling.

The contribution of this thesis can be categorised based on the typical steps of data handling lifecycle (i.e., provision, integration, exploitation) and summarised as follows:

- Provision

  - An OGD classification scheme that provides an understanding of the domain. The classification scheme is based on relevant literature and enables identifying, analysing and classifying OGD initiatives. The classification scheme comprises two main dimensions. The first dimension refers to the technological approach followed for making data available on the Web, while the second refers to the organisational approach followed for providing OGD. We believe that having a classification scheme allows a deeper understanding of initiatives and therefore the domain as a whole.
  - An OGD stage model that aims at (a) providing a roadmap for open government data re-use and (b) enabling evaluation of relevant initiatives' sophistication. The proposed model has two main dimensions, namely organisational & technological complexity and added value for data consumers.
  - A process model that describe the lifecycle of multi-dimensional OGD.

- Integration
  - Architectures and implementations for integrating OGD and social media data on the Linked Data Web.
  - A theoretical framework for integrating multi-dimensional OGD. This theoretical framework defines (a) binary relations that link two multi-dimensional data cubes that are compatible to integrate, and (b) operators that map from these two cubes to a new expanded one.
  - An analysis of the challenges that hamper the integration of multi-dimensional OGD using linked data paradigm.

- Exploitation in Data Analytics
  - A set of software tools that enable performing online analytical processing (OLAP) analytics on top of multiple datasets across the Linked Data Web. The tools were used to explore OGD from the Flemish government in Belgium and the Scottish government in the UK
  - A study to demonstrate the end-user value of the linked OGD analytics approach based on a case study that is related to the general elections of the UK using data from data.gov.uk, the official UK's OGD portal. The use case revealed that there is a significant relationship between the probability

one of the two main political parties (i.e. Labour Party and Conservative Party) to win in a UK constituency and the unemployment rate in the same constituency.

  - A process model that enables designing scientifically rigorous studies for exploiting social media data in predictive analytics. The model was used to design a case for predicting the winner of 2010 UK elections utilising linked open data to enrich Twitter data. The results suggested that the predictive power is weak for parties characterised by fluctuating voting intention over the last month before the election. We, however, extensively discussed each step of the approach to emphasise on the details that could affect the prediction accuracy.

- Access

  - An access control framework that capitalises on linked data paradigm to enable controlling access across distributed sources with diverse access constraints. The framework consists of (a) three linked data models, namely the LiMDAC metadata model for describing aggregated multi-dimensional data, the LiMDAC user profile model for describing data consumers, and the LiMDAC access policy model, and (b) an architecture that exploits and orchestrates the three models to enable controlling access to multi-dimensional data. From a technological perspective, the framework is validated using a proof-of-concept platform that is developed for that purpose using research multi-dimensional data.

7.2 Directions for Future Research

This thesis sets the stage for future research towards various directions. In this subsection we briefly describe those directions that we have already defined and we are working on:

7.2.1 Open Statistics

The most important direction for future research is related to further study and analyse open statistical data that can be modelled as multi-dimensional data cubes. Our long term vision is to enable performing statistical rigorous analytics on top of multiple open statistical datasets coming from disparate sources in an easy manner (Kalampokis et al., 2016b). Important activities towards this directions include:

- Identify best practices for publishing data cubes in the linked data Web. Towards this end we will capitalise on the analysis presented in 4.3.2.3 and we will employ a group of experts that have been involved in relevant activities across the globe. The adoption of the resulted set of best practices in future endeavours will ensure interoperability among linked data cubes publishers and thus will facilitate data cubes integration in the linked data Web.
- Identify and describe in detail more ways for combining data cubes on the Web, other than the one presented in 4.3.
- Develop more tools that will enable performing data analytics on top of multiple linked data cubes following various approaches ranging from panel data to statistical learning. These tools will adopt the approach followed by the OLAP Browser presented in 5.3. This means that the tools will be able to (a) identify compatible linked data cubes for a particular type of statistical analysis, (b) perform the analysis, and (c) present the results to the users. This will include performing analytics on a data store an automatically identifying relationships between variables that are described in the data.

7.2.2 Exploiting Linked Data Cubes in Public Service Co-Production

An other directions of future research is to apply the results of the thesis in a real-world environment in order to solve specific needs of the society. Towards this end, we will specify and realise an Innovation Ecosystem for Linked Open Statistical Data that is structured around societal needs and enables the co-production of effective and efficient data-driven public services that foster innovation and creativity in society and enterprises and thus stimulate economic growth in Europe. This requires (a) the modernisation of

public sector and (b) the collaboration of public sector, citizens and enterprises in order to translate societal needs into successful data-driven public services that exploit Linked Open Statistical Data. This ecosystem needs to be defined based on social, economical, technical, legal, political, institutional, and operational dimensions.

In this context we will perform the following research activities:

- Identify the challenges and needs (regarding legal, political, institutional, social, and technical issues) in opening-up and exploiting Linked Open Statistical Data (LOSD) for the co-production of innovative data-driven services. In this thesis we focused mainly on technical issues of OGD and linked data but in the future we will also study other challenges as well.
- Create a framework comprising processes, policies, and data infrastructure architecture that will specify a user-centric LOSD Innovation Ecosystem and will orchestrate the collaboration of society and public administration for opening up and exploiting LOSD in a way that will address all relevant challenges and facilitate the co-production of innovative data-driven services
- Develop open source and commercial ICT tools that will support the framework and enable public authorities to open up LOSD, and public administration and society to exploit this data in order to co-produce innovative services.
- Demonstrate the capability of the framework and the ICT tools in real world settings.

The work in this direction will be performed in the course of the research project "OpenGovIntelligence" that will be funded by the EU under Horizon 2020 program from 2016 until 2019.

7.2.3 Statistical models into the Linked Data Web

A very important direction for future research in the area is related to the idea that statistical models should be also opened up. An introduction to this idea has been presented in Kalampokis et al. (2013a) and O'reilly's Radar[1].

[1] http://radar.oreilly.com/2014/11/we-need-open-models-not-just-open-data.html

Different models could present controversial results in the same problem area and for the same variables depending on the statistical methods and/or the data that have been employed. For example, Chiricos (1987) reviewed 68 studies about the relationship between crime and the unemployment rate and he found that only less than half of these studies have found positive significant effects of the unemployment on crime rates. In addition, Kalampokis et al. (2013b) reviewed 52 empirical predictive models that employ predictors related to Social Media. They identified that the predictive power of a model is directly related to the predictors, the statistical method, the datasets and the evaluation method that have been selected. Thus, in order to better understand a problem we need to be able to discover and analyse various models that share common characteristics.

In addition, statistical models that have been developed based on a specific dataset can indeed be reused in another case. For example, a model developed for predicting sales based on data from Company X could be efficiently reused with data from Company Z. Moreover, a model predicting sales using a specific data mining method can be reused as a baseline for another model that uses a different method.

Publishing descriptions of statistical models on the Web following the Linked Data principles could have the following benefits:

1. Discovery of variables that a predictive relationship between them have been suggested by an empirical model. For example, it will be possible to discover that X number of models show a predictive relationship between product sales and advertising budget while Z number of models show a negative or no relationship between them.

2. Discovery of all predictor variables that are connected to product sales through successful empirical predictive models.

3. Discovery of statistical or data mining methods that have been used to identify relationships between variables. For example, most of the models that are able to accurately predict product sales from advertising budget have used linear regression methods.

4. Discovery of datasets that have been used to identify predictive relationships between variables. For example, models that show a strong predictive relationship

between product sales and advertising budget have employed data from the U.S. in the period between 1975 and 2004.

5. Discovery of a specific predictive model that shows a relationship between variables based on aspects such as its creator, the affiliation of the creator, the journal that the results have been published in, etc.

6. Discovery of new datasets in order to reuse existing models. For example, identification of datasets in Europe from the last ten years in order to reuse a predictive model produced with data from the U.S.

7. Discovery of predictive models that could be used as baseline models in building new more accurate predictive models.

These benefits will be achieved only if a vocabulary to model predictive models as RDF will be specified and Linked Data descriptions of predictive models will be published at a wide range. We believe that the adoption of such vocabulary could create new potentials beyond cross-platforms reuse of models. In particular, the vocabulary will enable (a) easy discovery and reuse of appropriate models at a Web Scale and (b) creation of more accurate models exploiting connections of models to other models, datasets and other resources on the Web.

A preliminary work towards this direction has been included in Appendix E, where RDF Linked Statistical Models *(LIMO)* vocabulary is described. *(LIMO)* allows for the description of statistical and data mining models in RDF and thus enables the incorporation of these models on the Linked Data Web and linking to others resources such as datasets, organisations, people and articles. However, more work towards this direction is needed.

Appendix A

Impact of the Thesis

Parts of this thesis have been published or have been submitted for publication in international refereed journals and conferences. In this appendix the references of these publications are presented along with citations to these publications only from journals that are indexed by the Web of Science (WoS) and thus an Impact Factor (IF) has been assigned to them.

Journal Articles

1. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (2016) "ICT Tools for Creating, Expanding, and Exploiting Statistical Linked Open Data", *Statistical Journal of the IAOS* [accepted for publication] (Indexed in Scopus)

2. Eleni Kamateri, Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (2014) "The Linked Medical Data Access Control Framework", *Journal of Biomedical Informatics*, Vol.50, pp. 213-225 (Indexed in WoS, **2014 IF: 2.194**) http://dx.doi.org/10.1016/j.jbi.2014.03.002

3. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (2013) "Understanding the Predictive Power of Social Media", *Internet Research*, Vol.23, No.5, pp. 544-559 (Indexed in WoS, **2013 IF: 1.638**) http://dx.doi.org/10.1108/IntR-06-2012-0114

4. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (2011) "A Classification Scheme for Open Government Data: Towards Linking Decentralized Data", *International Journal of Web Engineering and Technology*, Vol. 6, No. 3, pp.266-285. (Indexed in Scopus) http://dx.doi.org/10.1504/IJWET.2011.040725

5. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (XXX) "Expanding Data Cubes for Enhanced Analytics on the Web of Linked Data", *IEEE Transactions on Data and Knowledge Engineering* [submitted for review] (Indexed in WoS, **2014 IF: 2.067**)

6. Evangelos Kalampokis, Areti Karamanou, Efthimios Tambouris, and Konstantinos Tarabanis (XXX) "What Can Twitter and Linked Open Data Reveal about Elections Results? The Case of UK Election, 2010", *Journal of Intelligent Information Systems* [submitted for review] (Indexed in WoS, **2014 IF: 0.886**)

International Conference and Workshop papers

7. Evangelos Kalampokis, Efthimios Tambouris, Areti Karamanou, Konstantinos Tarabanis (2016) "Open Statistics: The Rise of a new Era for Open Data?, EGOV2016, LNCS, Springer [accepted for publication]

8. Evangelos Kalampokis, Bill Roberts, Areti Karamanou, Efthimios Tambouris, Konstantinos Tarabanis (2015) "Challenges on Developing Tools for Exploiting Linked Open Data Cubes" SemStats2015 in conjuction with the 14th International Semantic Web Conference (ISWC2015), 11-15 October 2015, Bethlehem, Pennsylvania, USA, CEUR-WS Vol.1551. http://ceur-ws.org/Vol-1551/article-07.pdf

9. Efthimios Tambouris, Evangelos Kalampokis, Konstantinos Tarabanis (2015) "Processing Linked Open Data Cubes", E. Tambouris, M. Janssen, H. J. Scholl, M. Wimmer, K. Tarabanis, M. Gascó, B. Klievink, I. Lindgren, and P. Parycek (Eds.): EGOV2015, LNCS 9248, pp.130-143, Springer. http://dx.doi.org/10.1007/978-3-319-22479-4_10

10. Efthimios Tambouris, Evangelos Kalampokis, Konstantinos Tarabanis (2015) "Create, Expand, and Exploit Linked Open Statistical Data", Electronic Government and Electronic Participation, Joint Proceedings of Ongoing Research, PhD

Papers, Posters and Workshops of IFIP EGOV and ePart 2015, pp.355-356, IOS Press. http://dx.doi.org/10.3233/978-1-61499-570-8-355

11. Efthimios Tambouris, Evangelos Kalampokis, Konstantinos Tarabanis (2015) "ICT Tools for statistical linked open data: The OpenCube toolkit", Proc. of the New Techniques and Technologies for Statistics Conference (NTTS2015), Brussels, Belgium, 10-12 March 2015. http://dx.doi.org/10.2901/EUROSTAT.C2015.001

12. Evangelos Kalampokis, Areti Karamanou, Andriy Nikolov, Peter Haase, Richard Cyganiak, Bill Roberts, Paul Hermans, Efthimios Tambouris, Konstantinos Tarabanis (2014) "Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services", Proc. of the 2nd International Workshop on Semantic Statistics (SemStats2014) in conjunction with the 13th International Semantic Web Confeerence (ISWC2014), 19-23 October 2014, Riva del Garda, Italy, CEUR-WS proceedings.

13. Evangelos Kalampokis, Andriy Nikolov, Peter Haase, Richard Cyganiak, Arkadiusz Stasiewicz, Areti Karamanou, Maria Zotou, Dimitris Zeginis, Efthimios Tambouris, Konstantinos Tarabanis (2014) "Exploiting Linked Data Cubes with OpenCube Toolkit", Proc. of the ISWC 2014 Posters and Demos Track a track within 13th International Semantic Web Confeerence (ISWC2014), 19-23 October 2014, Riva del Garda, Italy, CEUR-WS Vol.1272 http://ceur-ws.org/Vol-1272/paper_109.pdf

14. Evangelos Kalampokis, Areti Karamanou, Efthimios Tambouris, and Konstantinos Tarabanis (2013) "Towards a Vocabulary for Incorporating Predictive Models into the Linked Data Web", Proc. of the 1st International Workshop on Semantic Statistics (SemStats2013) in conjunction with the 12th International Semantic Web Conference (ISWC2013), 21-25 October 2013, Sydney, Australia, CEUR-WS proceedings.

15. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (2013) "Linked Open Government Data Analytics", M.A. Wimmer, M. Janssen, and H.J. Scholl (Eds.): EGOV 2013, LNCS 8074, pp. 99-110. Springer. http://dx.doi.org/10.1007/978-3-642-40358-3_9

16. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (2013) "On Publishing Linked Open Government Data", Proc. of the 17th Panhellenic Conference on Informatics (PCI 2013), Thessaloniki, Greece, pp.25-32, ACM. http://dx.doi.org/10.1145/2491845.2491869

17. Evangelos Kalampokis, Michael Hausenblas, and Konstantinos Tarabanis (2011) "Combining Social and Government Open Data for Participatory Decision-Making", E. Tambouris, A. Macintosh and H. de Bruijn (Eds): ePart2011, LNCS 6847, pp. 36-47, Springer. http://dx.doi.org/10.1007/978-3-642-23333-3_4

18. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis (2011) "Open Government Data: A Stage Model", M. Janssen et al. (Eds): EGOV2011, LNCS 6846, pp. 235-246, Springer. http://dx.doi.org/10.1504/10.1007/978-3-642-22878-0_20

Citations to these publications from papers published in journals that are indexed by WoS

2016

1. Wang, H.-J., Lo, J. (2016) Adoption of open government data among government agencies *Government Information Quarterly* (Indexed in WoS, **2014 IF:2.321**) http://dx.doi.org/10.1016/j.giq.2015.11.004

2. Wen Y.-F., Hung K.-Y., Hwang, Y.-T., Lin, Y.-S. F. (2016) Sports lottery game prediction system development and evaluation on social networks *Internet Research*, Vol. 26, No. 3, pp. 758-788 (Indexed in WoS, **2014 IF:1.661**) http://dx.doi.org/10.1108/IntR-05-2014-0139

3. Hilbert, M. (2016) Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, Vol. 34, No. 1, pp. 135-174 (Indexed in WoS, **2014 IF:1.024**) http://dx.doi.org/10.1111/dpr.12142

4. Charalabidis, Y., Alexopoulos, C., Loukis, E. (2016) A Taxonomy of Open Government Data Research Areas and Topics *Journal of Organizational Computing and Electronic Commerce*, (Indexed in WoS, **2014 IF:0.879**) http://dx.doi.org/10.1080/10919392.2015.1124720

5. Snelson, C. L. (2016) Qualitative and Mixed Methods Social Media Research: A Review of the Literature *International Journal of Qualitative Methods*, Vol. 15, No. 1, (Indexed in WoS, **2014 IF:0.494**) http://dx.doi.org/10.1177/1609406915624574

2015

6. Chenyan Xu, Yang Yu, and Chun-Keung Hoi (2015) Hidden in-game intelligence in NBA players' tweets. *Communications of the ACM*, Vo. 58, No. 11, pp. 80-89 (Indexed in WoS, **2014 IF:3.621**) http://dx.doi.org/10.1145/2735625

7. Andriotis, P., Oikonomou, G., Tryfonas, T., Li, S., (2015) Highlighting Relationships of a Smartphone's Social Ecosystem in Potentially Large Investigations, *IEEE Transactions on Cybernetics*, (Indexed in WoS, **2014 IF:3.469**) http://dx.doi.org/10.1109/TCYB.2015.2454733

8. Chae, B. (2015) Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research, *International Journal of Production Economics* Vol. 165, pp. 247–259 (Indexed in WoS, **2014 IF:2.752**) http://dx.doi.org/doi:10.1016/j.ijpe.2014.12.037

9. Yang Yu, Xiao Wang (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. *Computers in Human Behavior*, Vol. 48, pp. 392-400 (Indexed in WoS, **2014 IF:2.694**) http://dx.doi.org/10.1016/j.chb.2015.01.075

10. Attard, J., Orlandi, F., Scerri, S., Auer, S. (2015) A systematic review of open government data initiatives, *Government Information Quarterly*, Vol. 32, No. 4, pp. 399-418, (Indexed in WoS, **2014 IF: 2.321**) http://dx.doi.org/10.1016/j.giq.2015.07.006

11. Gonzalez-Zapata, F., Heeks, R. (2015) The multiple meanings of open government data: Understanding different stakeholders and their perspectives, *Government Information Quarterly*, Vol. 32, No. 4, pp. 441-452, (Indexed in WoS, **2014 IF: 2.321**) http://dx.doi.org/10.1016/j.giq.2015.09.001

12. Stamati, T., Papadopoulos, T., Anagnostopoulos, D. (2015) Social media for openness and accountability in the public sector: Cases in the Greek context, *Government Information Quarterly*, Vol. 32, No. 1, pp. 12-29, (Indexed in WoS, **2014 IF: 2.321**) http://dx.doi.org/10.1016/j.giq.2014.11.004

13. Leroux, H., Lefort, L. (2015) Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies, *Journal of Biomedical Semantics*, Vol. 6, No. 16, (Indexed in WoS, **2014 IF: 2.262**) http://dx.doi.org/10.1186/s13326-015-0012-6

14. Luarn, P. Chiu, Y.P. (2015) Key variables to predict tie strength on social network sites, *Internet Research*, Vol. 25, No. 2, pp.218 - 238 (Indexed in WoS, **2014 IF: 1.661**) http://dx.doi.org/10.1108/IntR-11-2013-0231

15. Duffett, R. G. (2015) Facebook advertising's influence on intention-to-purchase and purchase amongst Millennials, *Internet Research*, Vol. 25, No. 4, pp.498 - 526 (Indexed in WoS, **2014 IF: 1.661**) http://dx.doi.org/10.1108/IntR-01-2014-0020

16. Liljander, V., Gummerus, J., Söderlund, M. (2015) Young consumers' responses to suspected covert and overt blog marketing, *Internet Research*, Vol. 25, No. 4, pp.610 - 632 (Indexed in WoS, **2014 IF: 1.661**) http://dx.doi.org/10.1108/IntR-02-2014-0041

17. Lipizzi, C., Iandoli, L., Emmanuel Ramirez Marquez, J. E., (2015) Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams, *International Journal of Information Management*, Vol. 35, No. 4, pp. 490-503, (Indexed in WoS, **2014 IF: 1.550**) http://dx.doi.org/10.1016/j.ijinfomgt.2015.04.001

18. Martín, A. S., De Rosario, A. H., Pérez, M. D. C. C. (2015). An International Analysis of the Quality of Open Government Data Portals. *Social Science Computer Review*, [in press] (Indexed in WoS, **2014 IF: 1.364**) http://dx.doi.org/10.1177/0894439315585734

19. Nam, T. (2015). Challenges and Concerns of Open Government: A Case of Government 3.0 in Korea *Social Science Computer Review*, Vol. 33, No. 5,

pp. 556-570 (Indexed in WoS, **2014 IF: 1.364**) http://dx.doi.org/10.1177/0894439314560848

20. Yang T.-M., Lo, J., Shiang, J. (2015) To open or not to open? Determinants of open government data. *Journal of Information Science*, [in press] (Indexed in WoS, **2014 IF: 1.158**) http://dx.doi.org/10.1177/0165551515586715

21. Hossain, L., Hassan, M. R. Wigand, R. T. (2015) Resilient Information Networks for Coordination of Foodborne Disease Outbreaks, *Disaster Medicine and Public Health Preparedness*, Vol. 9, No. 2, pp. 186-198 **2014 IF: 1.142**) http://dx.doi.org/10.1017/dmp.2014.161

22. Wirtz, B., Piehler, R., Thomas, M.-J. (2015) Resistance of Public Personnel to Open Government: A cognitive theory view of implementation barriers towards open government data *Public Management Review* [in press] **2014 IF: 1.027**) http://dx.doi.org/10.1080/14719037.2015.1103889

23. Jiayin Pei , Guang Yu , Xianyun Tian , Maureen Renee Donnelley (2015) A new method for early detection of mass concern about public health issues, *Journal of Risk Research* [in press] **2014 IF: 0.935**) http://dx.doi.org/10.1080/13669877.2015.1100655

24. Hossain, M., Dwivedi, Y., Rana, N. (2015) State of the Art in Open Data Research: Insights from Existing Literature and a Research Agenda *Journal of Organizational Computing and Electronic Commerce*, Indexed in WoS, **2014 IF:0.879**) http://dx.doi.org/10.1080/10919392.2015.1124007

25. Sandoval-Almazan, R., Gil-Garcia, R.J. (2015) Towards an Integrative Assessment of Open Government: Proposing Conceptual Lenses and Practical Components *Journal of Organizational Computing and Electronic Commerce*, Indexed in WoS, **2014 IF:0.879**) http://dx.doi.org/10.1080/10919392.2015.1125190

26. Sivarajaha, U. Weerakkody, V. Waller, P. Lee, H. Irani, Z. Choi, Y. Morgan, R. Glikman, Y. (2015) The Role of e-Participation and Open Data in Evidence-Based Policy Decision Making in Local Government *Journal of Organizational Computing and Electronic Commerce*, Indexed in WoS, **2014 IF:0.879**) http://dx.doi.org/10.1080/10919392.2015.1125171

2014

27. Lausch, A., Schmidt, A. and Tischendorf, L. (2014). Data mining and linked open data - New perspectives for data analysis in environmental research. *Ecological Modelling*, Vol. 295, pp. 5-17 (Indexed in WoS, **2014 IF:2.321**) http://dx.doi.org/10.1016/j.ecolmodel.2014.09.018

28. Gerber, M. S. (2014). Predicting Crime Using Twitter and Kernel Density Estimation. *Decision Support Systems* Vol. 61, pp. 115-125 (Indexed in WoS, **2014 IF: 2.313**) http://dx.doi.org/10.1016/j.dss.2014.02.003

29. Whitmore, A. (2014). Using open government data to predict war: A case study of data and schema challenges. *Government Information Quarterly* Vol. 31, No. 4, pp. 622-630 (Indexed in WoS, **2014 IF: 2.321**) http://dx.doi.org/10.1016/j.giq.2014.04.003

30. Veljkovic, N., Bogdanovic-Dinic, S., Stoimenov, L. (2014) Benchmarking open government: An open data perspective, *Government Information Quarterly*, Vol. 31, No. 2, pp. 278-290, (Indexed in WoS, **2014 IF: 2.321**) http://dx.doi.org/10.1016/j.giq.2013.10.011.

31. Susha, I., Zuiderwijk, A., Janssen, M., Gronlund, A. (2014). Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned. *Social Science Computer Review*, Vol. 33, No. 5, pp. 613-630 (Indexed in WoS, **2014 IF: 1.364**) http://dx.doi.org/10.1177/0894439314560852

32. Ganapati, S., Reddick, C.G. (2014) The Use of ICT for Open Government in U. S. Municipalities, *Public Performance & Management Review*, Vol. 37, No. 3, pp. 365-387, (Indexed in WoS, **2014 IF: 0.641**) http://dx.doi.org/10.2753/PMR1530-9576370302

33. Solar, M., Daniels, F., Lopez, R., Meijueiro, L. (2014) A Model to Guide the Open Government Data Implementation in Public Agencies, *Journal of Universal Computer Science*, Vol. 20, No. 11, pp. 1564-1582, (Indexed in WoS, **2014 IF: 0.466**) http://dx.doi.org/10.3217/jucs-020-11-1564

2013

34. Maheshwari, D. and Janssen, M. (2013). Measurement and benchmarking foundations: Providing support to organizations in their development and growth using dashboards. *Government Information Quarterly*, Vol. 30, Supplement 1, pp. S83-S93. (Indexed in WoS, **2013 IF: 2.033**) http://dx.doi.org/10.1016/j.giq.2012.11.002

2012

35. Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W. and mc schraefel (2012). Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems*, Vol. 27, No.3, pp. 16-24. (Indexed in WoS, **2012 IF: 1.93**) http://dx.doi.org/10.1109/MIS.2012.23

36. Ding, L., Peristeras, V., Hausenblas, M., Linked Open Government Data [Guest editors' introduction], *IEEE Intelligent Systems*, Vol.27, No.3, pp.11-15, (Indexed in WoS, **2012 IF: 1.93**) http://dx.doi.org/10.1109/MIS.2012.56

37. Janssen, M., Charalabidis, Y. and Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, Vol. 29, No. 4, pp. 258-268. (Indexed in WoS, **2012 IF: 0.352**) http://dx.doi.org/10.1080/10580530.2012.716740

Appendix B

Evaluating Linked Data Tools for Handling Data Cubes

The exploitation of linked data in statistics requires specialised software tools that (a) are generic and thus applicable to all datasets that use the QB vocabulary, and (b) support each step of the linked data cube process. Therefore, existing linked data tools should be evaluated to determine their capability to fully support the process steps. In this Appendix, we present the results of the evaluation of nine widely-used open data, linked data, and statistical analysis tools, namely:

1. OpenRefine[1]
2. PoolParty[2]
3. CSVImport (LOD2 project)[3]
4. TabLinker[4]
5. SILK[5]
6. Pubby[6]

[1] http://openrefine.org
[2] http://www.poolparty.biz
[3] https://github.com/AKSW/csvimport.ontowiki
[4] https://github.com/Data2Semantics/TabLinker
[5] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/
[6] http://wifo5-03.informatik.uni-mannheim.de/pubby/

TABLE B.1: Evaluating the capacity of 9 tools to support the 8 steps of the process (Kalampokis et al., 2015)

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Step 1	F	N	P	P	N	N	N	N	N
Step 2	P	P	P	P	P	N	N	N	N
Step 3	N	N	P	P	N	P	N	N	N
Step 4	N	N	N	N	P	N	N	N	N
Step 5	N	N	N	N	N	N	N	N	N
Step 6	N	N	N	N	N	N	N	N	N
Step 7	N	N	N	N	N	N	P	P	P
Step 8	N	N	N	N	N	N	P	P	P

7. CubeViz (LOD2 project)[7]
8. SPARQL R[8]
9. RapidMiner (LOD)[9]

In Table B.1 the results of our analysis are presented. The horizontal axis presents the tools while the vertical the process steps. In each cell a letter indicates whether the tool (F)ully, (P)artially or N(ot) covers the functionality required by a step of the process. The analysis that we performed revealed that the following important functionalities are not currently supported by existing tools:

- Transform raw data to linked data cubes (as existing tools for RDF creation are difficult to use in the case of the QB vocabulary).
- Materialise cubes by computing aggregations across dimensions and hierarchies. This functionality is important for enabling OLAP browsing.
- Identify cubes with similar structure that could potential integrate.
- Create integrated views of multiple linked cubes on the Web. This will enable performing analytics on top of multiple cubes at a Web scale.

[7] http://cubeviz.aksw.org
[8] http://cran.r-project.org/web/packages/SPARQL/
[9] http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/

- Browse a linked data cube and perform advanced OLAP operations such as drill-down and roll-up.

Some tools, such as R and RapidMiner, with their extensions for importing RDF data can be also used for enabling performing data analytics on top of linked data cubes. We should, however, note that these generic RDF importers are difficult to use in the case of cubes because of the complexity that the QB vocabulary introduces. Our analysis revealed that linked data cube specific extensions are needed.

Appendix C

OGD Portals Providing Linked Data Cubes

At the moment, a number of linked data cubes are available on the Web. Some of them are official endeavours launched by government organisations that own the data. At this Appendix we describe the describe endeavours launched by the following governmental organisations:

- Scottish government in the UK
- UK Department for Communities and Local Government
- Italian National Institute of Statistics
- Irish Central Statistics Office
- Digital Agenda
- Flemish government in Belgium

The Scottish Government[1] provides the data behind their official statistics on "Neighborhood Statistics" as linked data (the site is currently in Beta version). They offer 129 linked data cubes categorised to 15 themes (e.g. housing, transport etc). The cubes comprise 17 distinct measures, 86 dimensions and 125 attributes. The offered cubes contain in average 657717 observations, while the geography dimension has 8475 distinct

[1]URL:http://statisticsbeta.com SPARQL endpoint:http://statisticsbeta.com/sparql

alues and the time dimension 139 distinct values. The geography and time has values at different granularity e.g. Parliamentary Constituencies, Council Areas etc. for the geography and 2002, 2002-Q1 etc for the time dimension.

The UK Department for Communities and Local Government (DCLG)[2] provides their official Linked Open Data of a selection of statistics. It provides a selection of statistics on a variety of themes including Local Government finance, housing and homelessness, wellbeing, deprivation, and the department's business plan as well as supporting geographical data. They offer 213 linked data cubes categorised to 14 themes (e.g. homelessness, societal wellbeing etc.). The offered cubes comprise 106 distinct measures, 94 dimensions and 129 attributes. The offered cubes contain in average 14437 observations, while the geography dimension has 89869 distinct values and the time dimension 110 distinct values. The geography and time has values at different granularity e.g. County, Region etc. for the geography and 2002, 2002-Q1 etc for the time dimension.

The Italian National Institute of Statistics (ISTAT)[3] makes available Italian Population and Housing Census 2011 as Linked Data. They offer 8 linked data cubes that comprise 8 distinct measures and 20 dimensions (they do not use attributes). The offered cubes contain in average 7060284 observations, while the geography dimension has 426725 distinct values at different granularity e.g. Region, Province etc. The time dimension does not exist since all data are only for 2011.

The Irish Central Statistics Office (CSO)[4] published census 2011 as linked data cubes providing a comprehensive picture of the social and living conditions of the people. They offer 682 linked data cubes that comprise 19 distinct measures and 50 dimensions (they do not use attributes). The offered cubes contain in average 5292 observations, while the geography dimension has 4806 distinct values at different granularity e.g. County, Electoral Division etc. The time dimension does not exist since all data are only for 2011. A peculiarity of the Irish Census data cubes is that they offer different data cubes for different geographical granularity instead of providing a single cube with all the values. For example they offer 12 cubes that measure the unemployment each one at different geographical granularity e.g. County, Electoral Division etc.

[2]URL:http://opendatacommunities.org SPARQL endpoint:http://opendatacommunities.org/sparql

[3]URL:http://datiopen.istat.it SPARQL endpoint:http://datiopen.istat.it/sparql

[4]URL:http://data.cso.ie SPARQL endpoint:http://data.cso.ie/query.html

The European Commission's Digital Agenda[5] provides its Scoreboard as linked data cubes. They offer 4 linked data cubes that comprise 7 distinct measures, 16 dimensions and 128 attributes. The offered cubes contain in average 145155 observations, while the geography dimension has 61 distinct values and the time dimension 77 distinct values. The geography and time has values at different granularity e.g. Greece, European Union 28 etc. for the geography and 2002, 2002-Q1, 2002-01 etc for the time dimension. A peculiarity of the Digital Agenda data cubes is that they use a "super-dimension" to embrace the values of dimensions other than time and location. This means that many cubes are conceptually integrated to the 4 offered cubes through the use of the super-dimension e.g. "Individuals who are born in non-EU country", "Individuals with high formal education" or "Unemployed".

The Flemish Government[6] makes their statistical link data cubes available with SKOS and XKOS hierarchies. The cubes have been produced through the OpenCube project with the official permission and contribution of the Flemish Government. Specifically, they offer 11 linked data cubes that comprise 27 distinct measures and 27 dimensions (they do not use attributes). The offered cubes contain in average 69687 observations, while the geography dimension has 589 distinct values and the time dimension 25 distinct values. The geography dimension has values at different granularity e.g. region, province, district etc.

[5]URL:http://digital-agenda-data.eu/data SPARQL endpoint:http://digital-agenda-data.eu/data/sparql

[6]URL:http://data.opendataforum.info SPARQL endpoint:http://188.166.18.242:8890/sparql

TABLE C.1: OGD portals providing linked data cubes

	Scottish	**DCLG**	**ISTAT**	**Irish CSO**	**Flemish**	**Digital Agenda**
Data	Neighbourhood Statistics	finance, well-being etc.	Italian Census 2011	Irish Census 2011	Flemish Gov. Datasets	Digital Agenda Scoreboard
Curator	Scottish Government	DCLG	ISTAT	Irish Central Statistics Office	Flemish Government	European Commission
Cubes	129	213	8	682	11	4
Measures	17	106	8	19	27	7
Dimensions	86	94	20	50	18	16
Attributes	125	129	0	0	0	128
Observations	84845456	3075142	56482270	3609306	766552	580620
Triples	901538411	126242629	800369986	20202132	7652149	4767031
GeoValues	8475	89869	426725	4806	589	61
TimeValues	139	110	-	-	25	77

Appendix D

UK Elections in Social Media

Below is presented a SPARQL query for identifying all the tweets that refer to the Conservative Party.

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT count(?tweet)
FROM <http://UKElections2010>
WHERE {
  ?tweet sioc:topic ?entity.
  {
     ?entity owl:sameAs dbpedia:Conservative_Party_(UK).
  }
UNION
  {
     ?entity owl:sameAs ?DBentity.
     ?DBentity (dbpedia-owl:party|dbpprop:party|dbpedia-owl:otherParty)
                    dbpedia:Conservative_Party_(UK).
  }

}
```

TABLE D.1: Hash tags used for classifying tweets about UK elections of 2010

Party	Party #	Positive #	Negative #
Conservatives	#cameron #tory #tories #conservatives #conservative #davidcameron	#imvotingconservative #voteforchange #votetory #torywin #imvotingtor #voteconservative #change	#imnotvotingconservative #toryfail #sameoldtories #keeptoriesout #anyonebutcameron #dontvotetory
Labours	#labour #gordonbrown #brown	#imvotinglabour #votelabour #labourwin #thankyoulabour	#labourdoorstep #labourfail #labourout #labourlost #labourisdead #imnotvotinglabour
Liberal Democrats	#libdems #libdem #clegg #nickclegg #nick_clegg	#imvotinglibdem #iagreewithnick #libdemmajority #votelibdem #imvotinglibdems #votelibdems #gonick	#nickcleggsfault #libdemfail
British National Party	#bnp		#bnpwatch #stopthebnp
UK Independence Party	#ukip		
Scottish National Party	#snp	#votesnp	
Pirate Party		#votepirate	
Green	#green #greenparty	#votegreen	#imnotvotinggreen

Appendix E

RDF Linked Statistical Models *(limo)* Vocabulary

In this Appendix we present the RDF Linked Statistical Models *(limo)* vocabulary that allows for the description of statistical and data mining models in the RDF model and thus enables the incorporation of these models on the Linked Data Web and linking to others resources such as datasets, organisations, people and articles. Although this is only a preliminary work that is considered as future work for this thesis, it indicates the importance of incorporating statistical models into the Linked Data Web.

In general, predictive analytics comprise predictive models designed for predicting new (or future) observations or scenarios as well as methods for evaluating the predictive power of a model Shmueli (2010). The outcome value for a new set of observation could be continuous (or quantitative) or categorical (or qualitative). In the former case the problem is ofter referred to as a regression problem while in the latter a classification problem. Predictive power refers to an empirical model's ability to predict new observations accurately. In contrast, explanatory power refers to the strength of association indicated by a statistical model.The predictive power of a model should be tasted based on out-of-sample data (e.g. cross-validation or a holdout sample) and with adequate predictive measures (e.g. RMSE, MAPE, PRESS etc.). A popular method to obtain out-of-sample data is to initially partition the data randomly, using one part (the training set) to fit the empirical model, and the other (the holdout set) to assess the model's predictive accuracy.

The vocabulary's main classes are depicted in Fig. E.1. Classes and properties from existing widely used vocabularies were reused whenever possible.

- `limo:Model` is the actual predictive model that is described by the vocabulary. The model has the following attributes:

  dct:title which is a name given to describe the model.

  dct:description for a descriptive comment about the model and its goals.

  dct:issued which defines the actual data that the model has been created.

  limo:modelType which describe the main categories of models that can be developed, namely classification, regression, clustering and dimensionReduction.

  limo:spatial is an attribute that describe the spatial dimension of the model. The spatial dimension of the model is derived from the actual data that have been employed. For example, a model could have `limo:spatial` U.S. in the case the data used for the development of the model comes from the U.S.

  limo:temporal is an attribute that describe the time period that the model covers. The time period of the model reflects the period that is described in the actual data that have been used for the development of the model.

  `limo:Model` is connected through `limo:data` property to a multi-dimensional data set i.e. a `qb:DataSet`. This dataset contains the actual data that have been used for the development of the model. As a result, the temporal and spatial dimension of the model could be also extracted from this dataset. In predictive analytics we have three different types of data, namely evaluation, validation and training data. So, *limo* includes three different sub-properties of the `limo:data` property, one for each of these three types of data.

  `limo:Model` is also connected through `limo:rawData` property to a `dctype: Dataset`. This dataset includes the raw data that have been used in the process of building the model. For example, this dataset could be a dump of raw tweets or a `dcat:Dataset` which thereafter was analysed in order to produce the actual data employed be the model.

  Moreover, the `limo:Model` can be connected to a different `limo:Model` through the `limo:baseline` property which explicitly denotes that the predictive power of a model has been evaluated against the power of another model.

The `limo:Model` can be also published in a scientific article or report. Hence we have included the `limo:publishedIn` property to express this relationship.

Finally, `limo:Model` is connected to a `foaf:Agent` through the `dct:creator` property. This property denotes the person or organisation that actually builds the model.

- `limo:Variable` represents the variables that are included in the predictive model. The Variable class includes the following attributes

  The dct:title denotes the actual name of the variable.

  The dct:description enables the inclusion of a small text in order to describe what the variable is about.

  The limo:variableType attributes denotes whether the variable is continuous, categorial or ordinal.

  The limo:usageType denotes whether the variable is the response of the model or one of the predictors.

  In addition, `limo:Variable` is categorised using the limo:theme property which connects the Variable to a `skos:Concept`

- `limo:Method` describes the statistical or data mining method used for creating the model. We assume that this class uses a set of predefined concepts such as linear regression, logistic regression, Markov models, support vector machine, random forests, neural networks etc. As a result, we assume that `limo:Method` is subclass of `skos:Concept`.

- `limo:Power` describes the predictive power of the model. The predictive power has the following attributes:

  limo:evaluationMethod is used to infer the predictive power of a model. The evaluation methods include out-of-sample evaluation with statistics such as Predicted Residual Sums of Squares, Root Mean Square Error or cross-validation techniques.

  limo:outcome is the actual value that the evaluation method produces.

- `limo:File` describes a file that can be imported in a particular platform such as R or SAS and execute the model. This could also be a PMML-XML file.

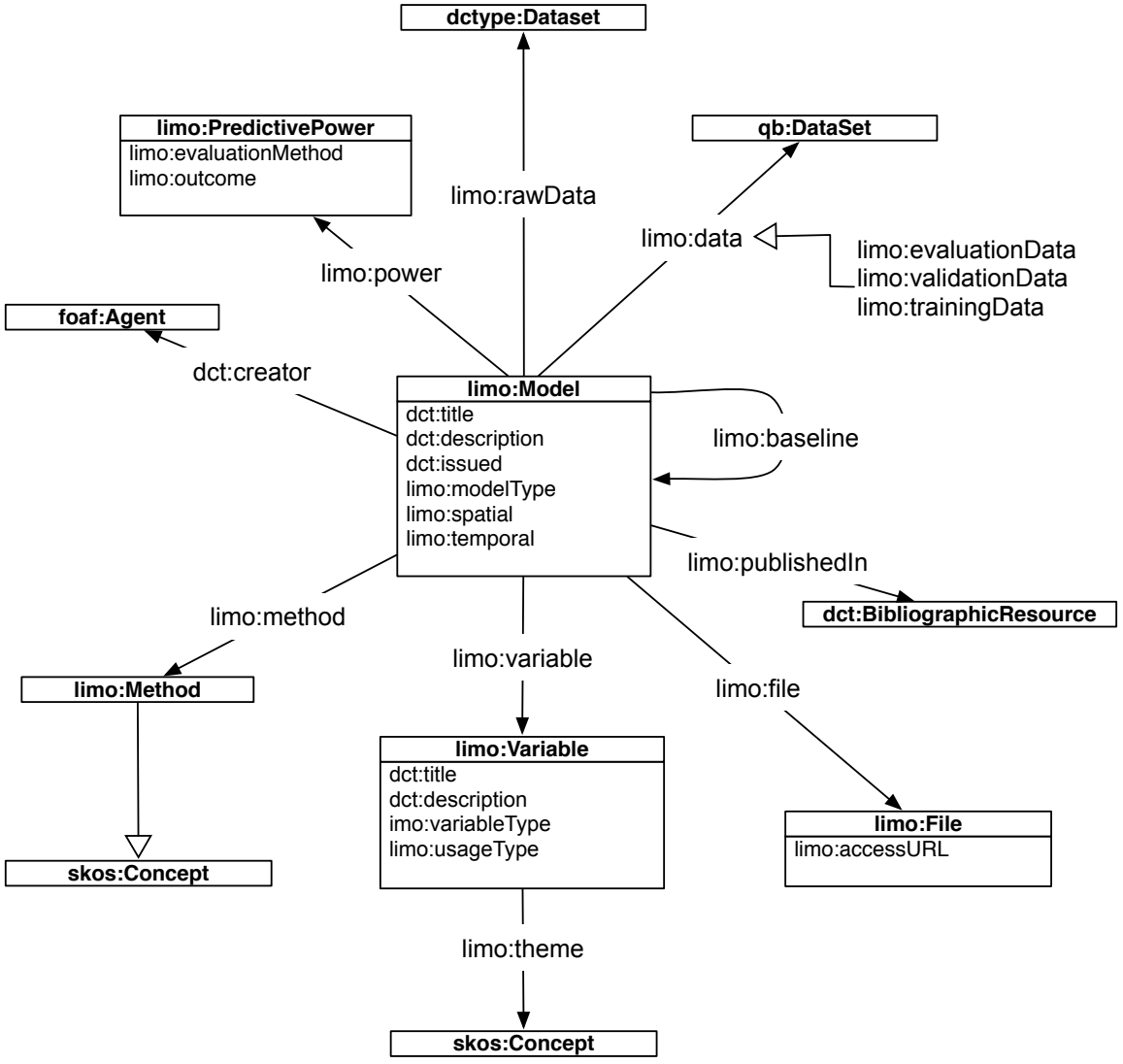


FIGURE E.1: The Linked Statistical Models vocabulary (Kalampokis et al., 2013a)

We should note that in this preliminary version of the vocabulary the execution of the model is possible through a PMML XML file. In the next version we aim at providing a more detailed description of the model in order to enable the execution of a model through its *limo* description. Full documentation of the *limo* vocabulary is available online[1].

In this section we present how *limo* vocabulary can be used in order (a) to describe a predictive model and (b) to enable the discovery of predictive models that address some requirements.

[1] http://purl.org/limo-ontology/limo

Below we present the *limo* description of the predictive model developed by Ginsberg et al. and presented in Ginsberg et al. (2009). This model aims at predicting influenza-like illness (ILI) physician visits from ILI-related queries. The models employs a linear regression method as well as data from Google and the US Centers for Disease Control and Prevention. The data is about nine regions of the United States between 2003 and 2008. The model was assessed using cross validation against out-of-sample data partitions and they obtained a mean correlation of 0.97.

Description of the predictive model presented in Ginsberg et al. (2009) with limo

```
eg:DDCILImodel a limo:Model;
    dct:title "CDC-ILI model"@en;
    limo:spatial [rdf:type dbpedia:United_States];
    limo:temporal
        [a dc:terms PeriodOfTime;
        limo:startDate "2003-09-28"^^xsd:date;
        limo:endDate "2008-05-11"^^xsd:date;];
    limo:modelType eg:regression;
    limo:variable eg:resp;
    limo:variable eg:pred;
    limo:method eg:linearregression;
    limo:power eg:CDCILIpower;
    limo:file eg:CDCILIfile;
    limo:rawData eg:CDCILIdataset;
    limo:evaluationData eg:CDCILIevaluationdata;
    limo:validationData eg:CDCILIvalidationdata;
    limo:trainingData eg:CDCILItrainingdata;
    dct:creator eg:ginsberg, eg:mohebbi, eg:patel, eg:brammer,
    eg:smolinski, eg:brilliant;
eg:resp a limo:Variable;
    limo:variableType eg:continuous;
    dct:description "Percentage of physician visits in which a
    patient presents with influenza-like symptoms in a region"@en;
    limo:usageType eg:response;
```

```
    limo:theme eg:ILIphysvisits.
eg:pred a limo:Variable;
    limo:variableType eg:continuous;
    dct:description "Probability that a random search query
    submitted from a region is ILI-related"@en;
    limo:usageType eg:predictor;
    limo:theme eg:ILIrandquery.
eg:CDCILIpower a limo:Power;
    limo:evaluationMethod eg:crossvalidation;
    limo:outcome 0.97.
eg:CDCILIdataset a dctype:DataSet;
    dct:resource <http://www.cdc.gov/flu/weekly>.
```

In addition, *limo* will enable the performance of queries across distributed description of predictive models. For example below we present a query answering the question "How many models exist that show relationship between the percentage of influenza-related physician visits and the probability that a random search query submitted from a region is influenza-related?".

A query for identifying models that predict influenza-like illnesses from search query data

```
SELECT (count( ?model ) as ?nmodels)
WHERE {
         {
            ?model limo:variable ?variable1;
                 limo:variable ?variable2.
            ?variable1 limo:usageType eg:response;
                 limo:theme eg:ILIphysvisits;
            ?variable2 limo:usageType eg:predictor;
                 limo:theme eg:ILIrandquery;
         } UNION
         {
            ?model limo:variable ?variable1;
                 limo:variable ?variable2.
```

```
            ?variable1 limo:usageType eg:predictor;
                 limo:theme eg: LIphysvisits.
            ?variable2 limo:usageType eg:response;
                 limo:theme eg: ILIrandquery.
          }

       }
```

Moreover, a query based on *limo* could unveil the variables that are predictors of influenza-related physician visits through empirical model(s) constructed by data regarding the U.S. The identification of these variables could enhance the process of building predictive model for influenza illnesses.

A query for identifying predictors of inluenza-like illnesses

```
SELECT ?variable
WHERE {
         ?model limo:variable ?variable1.
              limo:variable ?variable2.
              limo:spatial ?sp1.
         ?variable1 limo:usageType eg:predictor.
         ?variable2 limo:usageType eg:response;
              limo:theme eg:ILIphysvisits.
         ?sp1 rdf:type dbpedia:United_States.
       }
```

We believe that the adoption of the vocabulary could create new potentials beyond cross-platforms reuse of models. In particular, the vocabulary will enable (a) easy discovery and reuse of appropriate models at a Web Scale and (b) creation of more accurate models exploiting connections of models to other models, datasets and other resources on the Web.

Appendix F

Evaluation of NER tools in Twitter

In this Appendix we present the results of the evaluation of 8 Named Entity Recognition (NER) tools in Twitter data. To this end we again employ the UK election of 2010 use case and we collect data from Twitter that refer to this event. In order to collect the data we used #ge2010, #ukelection, #election2010 and #ge10 hashtags and we finally gathered 60.000 unique tweets from April 29, 2010 to May 6, 2010. For forming the gold standard 2000 tweets were randomly sampled from this collection and were manually annotated with 3 different Named Entity Types: Person, Organisation and Location. 2450 total entities were detected of which 950 were person entities, 1163 organisation entities and 337 were location entities.

NER methods can be classified into 2 main categories: Rule Based Methods and Machine Learning Methods. The former includes a set of rules that implements a specific grammar to identify and classify named entities while the latter is divided into 3 categories i.e. Supervised Learning (SL), Semi-Supervised Learning (SSL) and Unsupervised Learning (UL). In SL the classifier for NER is trained using a training data set that is annotated with named entities while in UL there is no annotated data. Existing NER tools can be divided into 2 further different categories i.e. a category comprising tools that can be implemented through Java libraries and a second one that consists of tools that provide APIs on the Web as web services.

Java based NER tools. Four tools are the most popular in this category: Apache OpenNLP's Name Finder[1], Lingpipe's NER tool [2] , Stanford NER[3] and GATE's ANNIE[4] . The first three are based on machine-learning methods. In particular Stan-ford NER implements linear chain Conditional Random Field (CRF) sequence models while Lingpipe's NER tool implements a first-order chain CRF. On the other hand, Apache OpenNLP's Name Finder uses a Maximum Entropy (ME) Model. They all provide pre-trained models that are trained on data sets from different domains like English News, Biomedical data, etc. They also provide the functionality to train the models on any other data set of one's choice. GATE's ANNIE is an Information Extraction system, which consists of various components for common NLP tasks. These components include tokeniser, sentence splitter, POS tagger, gazetteer, semantic tagger, orthomatcher and co-reference resolver. The semantic tagger does most of the work in NER and it is implemented as a set of rules written using Java Annotations Pattern Engine. These rules specify the patterns to be matched and annotations to be created as a result.

Web Services based NER tools. In this category 4 tools can be considered as the most important: AlchemyAPI[5] , DBpedia Spotlight[6] , Open Calais[7] and Zemanta[8]. Each service has different number of Named Entity Types (NET) it uses for classify-ing the entities detected in the text. For example, AlchemyAPI has around 100 differ-ent kinds of NET while Open Calais has 39. DBpedia Spotlight on the other hand uses DBpedia's class schema to generate the NET and Zemanta does not directly give the NET. Apart from doing NER like the tools in the first category these web services also provide URIs for the named entities detected in the text. These URIs are web resources from different sources, which also include data sets from the Linked Open Data Cloud. There are many other services that these tools offer such as suggesting tags, related articles and images, sentiment analysis, language detection, document categorisation, etc. but these services are not important for our work. Except DBPedia Spotlight all other 3 tools in this category have rate limits. For the free versions, AlchemyAPI has a limit of 30,000

[1] http://incubator.apache.org/opennlp/documentation/manual/opennlp.html#tools.namefind
[2] http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html
[3] http://nlp.stanford.edu/software/CRF-NER.shtml
[4] http://gate.ac.uk/sale/tao/splitch6.html#chap:annie
[5] http://www.alchemyapi.com/
[6] https://dbpedia-spotlight.github.io/demo/
[7] http://opencalais.com
[8] http://www.zemanta.com

calls per day, for Open Calais this limit is 50,000 calls per day, with a maximum of 4 calls per second, while for Zemanta the rate limit is 10,000 calls per day.

The first evaluation part includes the use of the 8 existing NER tools using their default settings in order to analyse 1000 tweets from the gold standard to identify and label the named entities that are present in the collected tweets. In order to further elaborate on the tools we perform a second round of evaluation by using in-domain training data. In this second round we use only Stanford NER and Apache OpenNLP as these two performed the best among the only 3 tools (the third one being Lingpipe) that enable modification of training data sets. These 2 tools were initially trained on 1000 manually annotated tweets (training data set) from the gold standard and then tested on the remaining 1000 tweets (test data set).

The entities that were identified within the boundary and labeled as the correct entity type were considered to be the correct entities (CE). The partially correct entities were the ones which were identified with boundary error (BE) but labeled correctly, the entities which were identified correctly within the boundaries but with a Labelling Error (LE) and those entities that had both BE and LE. Table F.1 gives an example of what exactly a CE, BE and LE are. For every tool we calculated the Recall (i.e. the ratio of the total number of CE identified to the total number of entities), the Precision (i.e. the ratio of the total number of CE identified to the total number of entities identified) and the F1 score (harmonic mean of Recall and Precision).

Although all these tools are capable of doing NER, they differ in many aspects. Due to this reason we establish common criteria on which all these tools can be com-pared. Below is a list of criteria, which we follow for our testing purpose:

- For any entity detected in the text, most of the times Zemanta gives more than one link to web resources to identify this entity. Hence, we can't be sure about which NET to take into account. In addition DBpedia suggest more than one Named Entity Type (NET) for any entity detected and secondly, many times this service does not suggest any Named Entity Type. So, in Zemanta and DBpedia Spotlight the detected entities will not be labeled with NET and thus there cannot be Labelling Errors taken into consideration. Therefore, to maintain consistent testing criteria between different tools we have considered partially correct entities

TABLE F.1: Correct Entities, Partially Correct Entities and Incorrect Entities

Tweet: I am donating my tweet to Sarah Carr and the Liberal Democrats to help them Get The Vote! #GE2010
Correct Result: I am donating my tweet to <PERSON>Sarah Carr</PERSON> and the <ORGANIZATION> Liberal Democrats </ORGANIZATION> to help them Get The Vote! #GE2010
Wrong Result: I am donating my tweet to <PERSON>Sarah</PERSON> Carr and the <ORGANIZATION>Liberal Democrats</ORGANIZATION> to help <PERSON>them</PERSON> Get The Vote! #GE2010 Correct Entities: Liberal Democrats Partially Correct Entities (Boundary Error): Sarah Incorrect Entities: them
Wrong Result: I am donating my tweet to <LOCATION>Sarah Carr </LOCATION> and the <LOCATION>Liberal</LOCATION> Democrats to help them Get The Vote! #GE2010 Partially Correct Entities (Labelling Error): Sarah Carr Partially Correct Entities (Boundary Error + Labelling Error): Liberal

as correct entities while calculating the F1 score. Nevertheless for the other 6 tools we have calculated the Boundary Error and Labelling Error.

- Different tool use different names for the NET. For example, AlchemyAPI uses City, Country, Continent, etc. whereas Stanford NER combines all these Named Entity Types into Location. Moreover the number of entity types detected is also different. For example, Open Calais detects 39 NET while for AlchemyAPI it is around 100. To even out such discrepancies 3 main categories of NET are considered, namely Person, Organisation and Location. So all other NET were considered as a sub category to these 3 main categories. For example, the NETs of AlchemyAPI City, Country, Continent, Geographic Feature, Region or County were all grouped together as Location. So any entity labeled with any of the NETs in subcategory will be an entity of the main category. Moreover those NETs that could not be classified into any of these 3 categories were ignored. For example, Open Calais has URL also as a NET, which was ignored for our testing.

Table F.2 depicts the overall results for all the 3 NETs while Tables F.3, F.4 and F.5 show the results for person, organisation and location NET respectively. All these tables indicate that Stanford NER present the best results. The in-domain trained Stanford NER showed significant improvement with the F1 score being increased by 16%.

Table F.2: Overall results

Tools	Recall	Precision	F1 score	TE	CE	BE	LE	BE+LE
Apache Open NLP	0.3479	0.7233	0.4698	571	334	33	38	8
GATE's ANNIE	0.3260	0.7979	0.4629	485	330	33	17	7
Lingpipe	0.5956	0.2583	0.3603	2737	419	135	97	56
Stanford NER	0.5805	0.8677	0.6956	794	595	29	45	20
Alchemy API	0.4575	0.8960	0.6058	606	501	33	6	8
DBpedia Spotlight	0.0446	0.4818	0.0816	112	52			
Open Calais	0.2384	0.9071	0.3776	312	232	20	27	4
Zemanta	0.5872	0.7012	0.6392	994	683			
Stanford NER (in-domain training)	0.7715	0.9686	0.8589	1018	936	30	14	6
Apache Open NLP (in-domain train.)	0.5282	0.7459	0.6185	905	607	34	31	3

Table F.3: Result for Named Entity Type Person

Tools	Recall	Precision	F1 score	TE	CE	BE	LE	BE+LE
Apache Open NLP	0.5495	0.8106	0.6550	301	206	23	14	1
GATE's ANNIE	0.4910	0.7814	0.6031	279	180	27	9	2
Lingpipe	0.5923	0.2447	0.3463	1075	173	58	15	17
Stanford NER	0.7838	0.9534	0.8603	365	320	15	13	
Alchemy API	0.7387	0.9162	0.8179	358	294	27	6	1
Open Calais	0.3288	0.9605	0.4899	152	122	12	10	2
Stanford NER (in-domain training)	0.7833	0.9724	0.8677	435	396	16	9	2
Apache Open NLP (in-domain train.)	0.3611	0.5891	0.4478	331	172	17	6	

Table F.4: Result for Named Entity Type Organization

Tools	Recall	Precision	F1 score	TE	CE	BE	LE	BE+LE
Apache Open NLP	0.1721	0.5650	0.2638	177	72	8	14	6
GATE's ANNIE	0.1583	0.7666	0.2624	120	75	4	8	5
Lingpipe	0.5250	0.2529	0.3496	1206	152	71	52	30
Stanford NER	0.3821	0.6033	0.4679	298	166	10	29	17
Alchemy API	0.2100	0.8652	0.3380	141	115	4	1	2
Open Calais	0.1394	0.7941	0.2372	102	61	6	12	2
Stanford NER (in-domain training)	0.7979	0.9763	0.8781	465	441	10	3	
Apache Open NLP (in-domain train.)	0.7153	0.8462	0.7753	481	381	7	17	2

Table F.5: Result for Named Entity Type Location

Tools	Recall	Precision	F1 score	TE	CE	BE	LE	BE+LE
Apache Open NLP	0.4259	0.7419	0.5411	93	56	2	10	1
GATE's ANNIE	0.4753	0.8953	0.6209	86	75	2		
Lingpipe	0.8580	0.3048	0.4498	456	94	6	30	9
Stanford NER	0.7284	0.9008	0.8055	131	109	4	2	3
Alchemy API	0.5802	0.8785	0.6988	107	92	2		
Open Calais	0.3333	0.9310	0.4909	58	49	2	3	
Stanford NER (in-domain training)	0.6450	0.9237	0.7596	118	99	4	5	1
Apache Open NLP (in-domain train.)	0.4320	0.7850	0.5574	93	54	10	8	1

Bibliography

F. Abel, J. L. D. Coi, N. Henze, A. W. Koesling, D. Krause, and D. Olmedilla. Enabling advanced and context-dependent access control in RDF stores. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.*, pages 1–14, 2007. doi: 10.1007/978-3-540-76298-0_1. URL http://dx.doi.org/10.1007/978-3-540-76298-0_1.

F. Abel, I. Celik, G.-J. Houben, and P. Siehndel. Leveraging the semantics of tweets for adaptive faceted search on twitter. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist, editors, *The Semantic Web – ISWC 2011*, volume 7031 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-25072-9. doi: 10.1007/978-3-642-25073-6_1. URL http://dx.doi.org/10.1007/978-3-642-25073-6_1.

A. Abello, O. Romero, T. Pedersen, R. Berlanga Llavori, V. Nebot, M. Aramburu, and A. Simitsis. Using semantic web technologies for exploratory olap: A survey, 2014. ISSN 1041-4347.

H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702 –707, april 2011. doi: 10.1109/INFCOMW.2011.5928903.

R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. In *Data Engineering, 1997. Proceedings. 13th International Conference on*, pages 232–243, Apr 1997. doi: 10.1109/ICDE.1997.581777.

B. M. Althouse, Y. Y. Ng, and D. A. T. Cummings. Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis*, 5(8):e1258, 08 2011.

doi: 10.1371/journal.pntd.0001258. URL http://dx.doi.org/10.1371%2Fjournal.pntd.0001258.

K. V. Andersen and H. Z. Henriksen. E-government maturity models: Extension of the layne and lee model. *Government Information Quarterly*, 23(2):236 – 248, 2006. ISSN 0740-624X. doi: http://dx.doi.org/10.1016/j.giq.2005.11.008. URL http://www.sciencedirect.com/science/article/pii/S0740624X05000973.

C. Anderson. The end of theory: The data deluge makes the scientific method obsolete, 2008. URL http://www.uvm.edu/~cmplxsys/wordpress/wp-content/uploads/reading-group/pdfs/2008/anderson2008.pdf.

N. Anderson and K. Edwards. Building a chain of trust: using policy and practice to enhance trustworthy clinical data discovery and sharing. In *Proceedings of the 2010 Workshop on Governance of Technology, Information and Policies*, pages 15–20. ACM, 2010.

N. R. Anderson, A. Abend, A. Mandel, E. M. Geraghty, D. Gabriel, R. Wynden, M. Kamerick, K. Anderson, J. Rainwater, and P. Tarczy-Hornoch. Implementation of a deidentified federated data network for population-based cohort discovery. *JAMIA*, 19(e1), 2012. doi: 10.1136/amiajnl-2011-000133. URL http://dx.doi.org/10.1136/amiajnl-2011-000133.

A. Antoniades, C. Georgousopoulos, N. Forgo, A. Aristodimou, F. Tozzi, P. Hasapis, K. Perakis, T. Bouras, D. Alexandrou, E. Kamateri, E. Panopoulou, K. Tarabanis, and C. Pattichis. Linked2safety: A secure linked data medical information space for semantically-interconnecting ehrs advancing patients' safety in medical research. *13th IEEE International Conference on BioInformatics and BioEngineering*, 0:517–522, 2012. doi: http://doi.ieeecomputersociety.org/10.1109/BIBE.2012.6399767.

W. Antweiler and M. Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2004.00662.x. URL http://dx.doi.org/10.1111/j.1540-6261.2004.00662.x.

S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence*

and Intelligent Agent Technology - Volume 01, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4191-4. doi: 10.1109/WI-IAT.2010.63. URL http://dx.doi.org/10.1109/WI-IAT.2010.63.

S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller. Triplify: Light-weight linked data publication from relational databases. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 621–630, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526793. URL http://doi.acm.org/10.1145/1526709.1526793.

S. Auer, J. Lehmann, A.-C. Ngonga Ngomo, and A. Zaveri. Introduction to linked data and its lifecycle on the web. In S. Rudolph, G. Gottlob, I. Horrocks, and F. van Harmelen, editors, *Reasoning Web. Semantic Technologies for Intelligent Data Access*, volume 8067 of *Lecture Notes in Computer Science*, pages 1–90. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-39783-7. doi: 10.1007/978-3-642-39784-4_1. URL http://dx.doi.org/10.1007/978-3-642-39784-4_1.

R. C. Barrows and P. D. Clayton. Privacy, Confidentiality, and Electronic Medical Records. *Journal of the American Medical Informatics Association*, 3(2):139–148, Mar. 1996. ISSN 1067-5027. doi: 10.1136/jamia.1996.96236282. URL http://dx.doi.org/10.1136/jamia.1996.96236282.

A. Begoyan. An overview of interoperability standards for electronic health records. In *Integrated Design and Process Technology, IDPT-2007*, 2007.

S.-M.-R. Beheshti, B. Benatallah, H. Motahari-Nezhad, and M. Allahbakhsh. A framework and a language for on-line analytical processing on graphs. In X. Wang, I. Cruz, A. Delis, and G. Huang, editors, *Web Information Systems Engineering - WISE 2012*, volume 7651 of *Lecture Notes in Computer Science*, pages 213–227. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-35062-7. doi: 10.1007/978-3-642-35063-4_16. URL http://dx.doi.org/10.1007/978-3-642-35063-4_16.

T. Berners-Lee. Putting governmet data online, 2009. URL http://www.w3.org/DesignIssues/GovData.html.

T. Berners-Lee. Design issues: Linked data, 2010. URL http://www.w3.org/DesignIssues/LinkedData.html.

C. Bizer and R. Cyganiak. D2r server-publishing relational databases on the semantic web. In *Poster at the 5th International Semantic Web Conference*, 2006.

C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009a.

C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. {DBpedia} - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165, 2009b. ISSN 1570-8268. doi: http://dx.doi.org/10.1016/j.websem.2009.07.002. URL http://www.sciencedirect.com/science/article/pii/S1570826809000225. The Web of Data.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011. ISSN 1877-7503. doi: http://dx.doi.org/10.1016/j.jocs.2010.12.007. URL http://www.sciencedirect.com/science/article/pii/S187775031100007X.

I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen, and I. Weber. Web search queries can predict stock market volumes. *PLoS ONE*, 7(7):e40014, 07 2012. doi: 10.1371/journal.pone.0040014. URL http://dx.doi.org/10.1371%2Fjournal.pone.0040014.

T. Bosch, R. Cyganiak, A. Gregory, and J. Wackerow. DDI-RDF discovery vocabulary: A metadata vocabulary for documenting research and survey data. In *Proceedings of the WWW2013 Workshop on Linked Data on the Web, Rio de Janeiro, Brazil, 14 May, 2013*, 2013. URL http://ceur-ws.org/Vol-996/papers/ldow2013-paper-12.pdf.

E. Bothos, D. Apostolou, and G. Mentzas. Using social media to predict future events with agent-based markets. *IEEE Intelligent Systems*, 25(6):50–58, 2010. ISSN 1541-1672. doi: http://doi.ieeecomputersociety.org/10.1109/MIS.2010.152.

J. Breen. At the dawn of e-government: the citizen as customer. *Government Finance Review*, 16(5):15–20, 2000.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 08 2001. doi: 10.1214/ss/1009213726. URL http://dx.doi.org/10.1214/ss/1009213726.

M. Burdon. Commercializing public sector information privacy and security concerns. *Technology and Society Magazine, IEEE*, 28(1):34–40, Spring 2009. ISSN 0278-0097. doi: 10.1109/MTS.2009.931860.

A. Burgun and O. Bodenreider. Accessing and integrating data and knowledge for biomedical research. *Yearbook of medical informatics*, pages 91–101, 2008. ISSN 0943-4747. URL http://view.ncbi.nlm.nih.gov/pubmed/18660883.

L. Cabibbo and R. Torlone. From a procedural to a visual query language for olap. In *Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on*, pages 74–83, Jul 1998. doi: 10.1109/SSDM.1998.688113.

K. Caine and R. Hanania. Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association : JAMIA*, 20(1):7–15, Jan. 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2012-001023. URL http://dx.doi.org/10.1136/amiajnl-2012-001023.

D. Calvanese, G. de Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. Data integration in data warehousing. *International Journal of Cooperative Information Systems*, 10(03):237–271, 2001. doi: 10.1142/S0218843001000345. URL http://www.worldscientific.com/doi/abs/10.1142/S0218843001000345.

D. Cameron. Letter to government departments on opening up, May 2010. URL http://webarchive.nationalarchives.gov.uk/20130109092234/http://www.number10.gov.uk/news/statements-and-articles/2010/05/letter-to-government-departments-on-opening-up-data-51204.

S. Capadisli, S. Auer, and A.-C. Ngonga Ngomo. Linked sdmx data. *Semantic Web*, 2013.

C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963500. URL http://doi.acm.org/10.1145/1963405.1963500.

S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. *SIGMOD Rec.*, 26(1):65–74, Mar. 1997. ISSN 0163-5808. doi: 10.1145/248603.248616.

C. Chen, X. Yan, F. Zhu, J. Han, and P. Yu. Graph OLAP: a multi-dimensional framework for graph data analysis. *Knowledge and Information Systems*, 21(1):41–63, 2009. ISSN 0219-1377. doi: 10.1007/s10115-009-0228-9. URL http://dx.doi.org/10.1007/s10115-009-0228-9.

Y. Chen, Q. Wang, and J. Xie. Online social interactions: A natural experiment on word of mouth versus observational learning. *Journal of Marketing Research*, 48(2): 238–254, 2013/01/17 2011. doi: 10.1509/jmkr.48.2.238. URL http://dx.doi.org/10.1509/jmkr.48.2.238.

J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2013/01/17 2006. doi: 10.1509/jmkr.43.3.345. URL http://dx.doi.org/10.1509/jmkr.43.3.345.

T. G. Chiricos. Rates of crime and unemployment: An analysis of aggregate research evidence. *Social Problems*, 34(2):187–212, 1987. ISSN 00377791, 15338533. URL http://www.jstor.org/stable/800715.

H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88:2–9, 2012. ISSN 1475-4932. doi: 10.1111/j.1475-4932.2012.00809.x. URL http://dx.doi.org/10.1111/j.1475-4932.2012.00809.x.

R. Chunara, J. R. Andrews, and J. S. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, 2012. doi: 10.4269/ajtmh.2012.11-0597. URL http://www.ajtmh.org/content/86/1/39.abstract.

E. Codd, S. Codd, and C. Salley. *Providing OLAP (On-line Analytical Processing) to User-analysts: An IT Mandate*. Codd & Associates, 1993. URL http://books.google.gr/books?id=pt0lGwAACAAJ.

D. Colazzo, F. Goasdoué, I. Manolescu, and A. Roatiş. RDF analytics: Lenses over semantic graphs. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 467–478, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2744-2. doi: 10.1145/2566486.2567982. URL http://doi.acm.org/10.1145/2566486.2567982.

T. U. Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research*, 41(D1):D43–D47, 2013. doi: 10.1093/nar/gks1068. URL http://dx.doi.org/10.1093/nar/gks1068.

C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh. Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7(2):596–615, 2010. ISSN 1660-4601. doi: 10.3390/ijerph7020596. URL http://www.mdpi.com/1660-4601/7/2/596.

L. Costabello, S. Villata, N. Delaforge, and F. Gandon. Linked data access goes mobile: Context-aware authorization for graph stores. In *WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012*, 2012. URL http://ceur-ws.org/Vol-937/ldow2012-paper-05.pdf.

D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. K. Yung, E. Birney, H. Hermjakob, P. D'Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database-Issue):691–697, 2011. doi: 10.1093/nar/gkq1018. URL http://dx.doi.org/10.1093/nar/gkq1018.

A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0217-3. doi: 10.1145/1964858.1964874. URL http://doi.acm.org/10.1145/1964858.1964874.

G. G. Curtin. Free the data!: E-governance for megaregions. *Public Works Management & Policy*, 14(3):307–326, 2010. doi: 10.1177/1087724X09359352. URL http://pwm.sagepub.com/content/14/3/307.abstract.

R. Cyganiak and D. Reynolds. The rdf data cube vocabulary: W3c recommendation. Technical report, W3C, January 2014.

Z. Da, J. Engelberg, and P. Gao. In search of attention. *The Journal of Finance*, 66 (5):1461–1499, 2011. ISSN 1540-6261. doi: 10.1111/j.1540-6261.2011.01679.x. URL http://dx.doi.org/10.1111/j.1540-6261.2011.01679.x.

A. Datta and H. Thomas. The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. *Decision Support Systems*, 27(3):289 – 301, 1999. ISSN 0167-9236. doi: http://dx.doi.org/10.1016/S0167-9236(99)00052-4. URL http://www.sciencedirect.com/science/article/pii/S0167923699000524.

M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Can blog communication dynamics be correlated with stock market activity? In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, HT '08, pages 55–60, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-985-2. doi: 10.1145/1379092.1379106. URL http://doi.acm.org/10.1145/1379092.1379106.

M. Dekkers, F. Polman, R. te Velde, and M. de Vries. Mepsir: Measuring european public sector information resources. Technical report, European Commission, DG Information Society, June 2006.

Deloitte. Market assessment of public sector information. Technical report, UK Department for Business Innovation & Skills, May 2013. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198905/bis-13-743-market-assessment-of-public-sector-information.pdf.

R. V. Dhopeshwarkar, L. M. Kern, H. C. O'Donnell, A. M. Edwards, and R. Kaushal. Health care consumers' preferences around health information exchange. *The Annals of Family Medicine*, 10(5):428–434, 2012.

L. Ding, V. Peristeras, and M. Hausenblas. Linked open government data [guest editors' introduction]. *Intelligent Systems, IEEE*, 27(3):11–15, May 2012. ISSN 1541-1672. doi: 10.1109/MIS.2012.56.

T. A. Drake, J. Braun, A. Marchevsky, I. S. Kohane, C. Fletcher, H. Chueh, B. Beckwith, D. Berkowicz, F. Kuo, Q. T. Zeng, et al. A system for sharing routine surgical pathology specimens across institutions: the shared pathology informatics network. *Human pathology*, 38(8):1212–1225, 2007.

W. Duan, B. Gu, and A. B. Whinston. Do online reviews matter? an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007 – 1016, 2008. ISSN 0167-9236. doi: 10.1016/j.dss.2008.04.001. URL http://www.sciencedirect.com/

science/article/pii/S0167923608000754. ¡ce:title¿Information Technology and Systems in the Internet-Era¡/ce:title¿.

P. Earle, D. Bowden, and M. Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012. ISSN 2037-416X. URL http://www.annalsofgeophysics.eu/index.php/annals/article/view/5364.

L. Etcheverry and A. Vaisman. Enhancing olap analysis with web cubes. In E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, editors, *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, pages 469–483. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-30283-1. doi: 10.1007/978-3-642-30284-8_38. URL http://dx.doi.org/10.1007/978-3-642-30284-8_38.

L. Etcheverry, A. Vaisman, and E. Zimányi. Modeling and querying data warehouses on the semantic web using qb4olap. In L. Bellatreche and M. Mohania, editors, *Data Warehousing and Knowledge Discovery*, volume 8646 of *Lecture Notes in Computer Science*, pages 45–56. Springer International Publishing, 2014. ISBN 978-3-319-10159-0. doi: 10.1007/978-3-319-10160-6_5. URL http://dx.doi.org/10.1007/978-3-319-10160-6_5.

M. Ettredge, J. Gerdes, and G. Karuga. Using web-based search data to predict macroeconomic statistics. *Commun. ACM*, 48(11):87–92, Nov. 2005. ISSN 0001-0782. doi: 10.1145/1096000.1096010. URL http://doi.acm.org/10.1145/1096000.1096010.

European Commission. Public sector information: A key resource for europe. green paper on public sector information in the information society. Technical Report COM(1998)585, European Commission, 1998.

European Commission. Commercial exploitation of europe's public sector information. Final Report, October 2000. Office for Official Publications of the European Communities. Luxembourg, October 2000.

European Commission. eeurope2002: Creating a eu framework for the exploitation of public sector information. Technical Report COM(2001) 607, European Commission, October 2001.

European Commission. Directive 2003/98/ec of the european parliament and of the council on the re-use of public sector information'. *Official Journal of the European Union*, 345:90–96, 2003.

European Commission. Re-use of public sector informaiton - review of directive 2003/98/ec. Technical Report COM(2009) 212, European Commission, May 2009.

European Commission. Open data: An engine for innovation, growth and transparent governance. Communication from the Commission, COM(2011) 882 final, December 2011.

A. M. Evans and A. Campos. Open government initiatives: Challenges of citizen participation. *Journal of Policy Analysis and Management*, 32(1):172–185, 2013.

J. L. Fernández-Alemán, I. C. Señor, P. Á. O. Lozoya, and A. Toval. Security and privacy in electronic health records: A systematic literature review. *Journal of Biomedical Informatics*, 46(3):541 – 562, 2013. ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2012.12.003. URL http://www.sciencedirect.com/science/article/pii/S1532046412001864.

T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1866696.1866709.

G. Flouris, I. Fundulaki, M. Michou, and G. Antoniou. Controlling access to RDF graphs. In *Future Internet - FIS 2010 - Third Future Internet Symposium, Berlin, Germany, September 20-22, 2010. Proceedings*, pages 107–117, 2010. doi: 10.1007/978-3-642-15877-3_12. URL http://dx.doi.org/10.1007/978-3-642-15877-3_12.

C. Forman, A. Ghose, and B. Wiesenfeld. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3):291–313, 2008. doi: 10.1287/isre.1080.0193. URL http://pubsonline.informs.org/doi/abs/10.1287/isre.1080.0193.

F. Franch. (wisdom of the crowds)2: 2010 uk election prediction with social media. *Journal of Information Technology & Politics*, 10(1):57–71, 2013. doi: 10.1080/19331681.2012.705080. URL http://www.tandfonline.com/doi/abs/10.1080/19331681.2012.705080.

D. Gayo-Avello. Don't turn social media into another 'literary digest' poll. *Commun. ACM*, 54(10):121–128, Oct. 2011. ISSN 0001-0782. doi: 10.1145/2001269.2001297. URL http://doi.acm.org/10.1145/2001269.2001297.

S. Ghasemi, W.-S. Luk, and N. Alrayes. M2RML: Multidimensional to RDF Mapping Language. In *Database and Expert Systems Applications (DEXA), 2014 25th International Workshop on*, pages 263–267, Sept 2014. doi: 10.1109/DEXA.2014.61.

A. Ghose and P. G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on*, 23(10):1498–1512, Oct 2011. ISSN 1041-4347. doi: 10.1109/TKDE.2010.188.

E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 211–220, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518736. URL http://doi.acm.org/10.1145/1518701.1518736.

E. Gilbert and K. Karahalios. Widespread worry and the stock market. In *ICWSM*, pages 59–65, 2010.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 02 2009. URL http://dx.doi.org/10.1038/nature07634.

S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 2010. doi: 10.1073/pnas.1005962107. URL http://www.pnas.org/content/early/2010/09/20/1005962107.abstract.

L. I. Gómez, S. A. Gómez, and A. A. Vaisman. A generic data model and query language for spatiotemporal olap cube analysis. In *Proceedings of the 15th International Conference on Extending Database Technology*, EDBT '12, pages 300–311, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0790-1. doi: 10.1145/2247596.2247632. URL http://doi.acm.org/10.1145/2247596.2247632.

W. Goossen, A. Goossen-Baremans, and M. van der Zel. Detailed Clinical Models: A Review. *Healthc Inform Res*, 16(4):201–214, Dec. 2010. URL http://synapse.koreamed.org/DOIx.php?id=10.4258/hir.2010.16.4.201&vmode=FULL.

J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery*, 1(1):29–53, 1997.

C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pages 27–37, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0245-6. doi: 10.1145/1866307.1866311. URL http://doi.acm.org/10.1145/1866307.1866311.

D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 78–87, New York, NY, USA, 2005. ACM. ISBN 1-59593-135-X. doi: 10.1145/1081870.1081883. URL http://doi.acm.org/10.1145/1081870.1081883.

M. B. Gurstein. Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 16(2), 2011.

G. Guzman. Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, 36(3):119–167, 2011.

P. Haase, M. Schmidt, and A. Schwarte. The information workbench as a self-service platform for linked data applications. Citeseer.

B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002520.

Q. Hardy. Better economic forecasts, from the cloud, 2012. URL http://bits.blogs.nytimes.com/2012/03/15/better-forecasts-from-the-cloud/?_r=0.

M. Hausenblas. Exploiting linked data to build web applications. *IEEE Internet Computing*, 13(4):68–73, July 2009. ISSN 1089-7801. doi: 10.1109/MIC.2009.79. URL http://dx.doi.org/10.1109/MIC.2009.79.

Y. He, H. Saif, Z. Wei, and K.-F. Wong. Quantising opinions for political tweets analysis. In E. L. R. Association, editor, *Eight International Conference on Language Resources and Evaluation*, pages 3901–3906, 2012.

T. Heath. How will we interact with the web of data? *Internet Computing, IEEE*, 12 (5):88–91, 2008.

N. Helbig, A. M. Cresswell, G. B. Burke, and L. Luna-Reyes. The dynamics of opening government data. *Center for Technology in Government.[Online]. Available: http://www. ctg. albany. edu/publications/reports/opendata*, 2012.

J. Helmich, J. Klímek, and M. Nečaský. Visualizing rdf data cubes using the linked data visualization model. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events*, Lecture Notes in Computer Science, pages 368–373. Springer International Publishing, 2014. ISBN 978-3-319-11954-0. doi: 10.1007/978-3-319-11955-7_50. URL http://dx.doi.org/10.1007/978-3-319-11955-7_50.

K. Höffner and J. Lehmann. Towards question answering on statistical linked data. In *Proceedings of the 10th International Conference on Semantic Systems*, SEM '14, pages 61–64, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2927-9. doi: 10.1145/2660517.2660521. URL http://doi.acm.org/10.1145/2660517.2660521.

L.-C. Huang, H.-C. Chu, C.-Y. Lien, C.-H. Hsiao, and T. Kao. Privacy preservation and information security protection for patients' portable electronic health records. *Comput. Biol. Med.*, 39(9):743–750, Sept. 2009. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2009.06.004. URL http://dx.doi.org/10.1016/j.compbiomed.2009.06.004.

A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009. doi: 10.1504/IJEM.2009.031564.

A. Hulth, G. Rydevik, and A. Linde. Web queries as a source for syndromic surveillance. *PLoS ONE*, 4(2):e4378, 02 2009. doi: 10.1371/journal.pone.0004378. URL http://dx.doi.org/10.1371%2Fjournal.pone.0004378.

B. Hyland and D. Wood. The joy of data - a cookbook for publishing linked government data on the web. In D. Wood, editor, *Linking Government Data*, pages 3–26. Springer

New York, 2011. ISBN 978-1-4614-1766-8. doi: 10.1007/978-1-4614-1767-5_1. URL http://dx.doi.org/10.1007/978-1-4614-1767-5_1.

W. H. Inmon. Building the data warehouse. *John wiley & so Golfarelli, M., Maio, D., & Rizzi, S.(1998). The dimensional fact model: Aconceptual model for data warehouses. International Journal of CooperativeInformation Systems*, 7:215–247, 2005.

A. Jain and C. Farkas. Secure resource description framework: an access control model. In *SACMAT 2006,11th ACM Symposium on Access Control Models and Technologies, Lake Tahoe, California, USA, June 7-9, 2006, Proceedings*, pages 121–129, 2006. doi: 10.1145/1133058.1133076. URL http://doi.acm.org/10.1145/1133058.1133076.

M. Janssen and A. Zuiderwijk. Open data and transformational government. In *TGov Conference, London*, 2012.

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

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-848-0. doi: 10.1145/1348549.1348556. URL http://doi.acm.org/10.1145/1348549.1348556.

P. Jensen, L. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews. Genetics*, 13(6):395–405, 2012. ISSN 1471-0056. doi: 10.1038/nrg3208.

X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The wisdom of social multimedia: Using flickr for prediction and forecast. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1235–1244, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874196. URL http://doi.acm.org/10.1145/1873951.1874196.

Z. Jourdan, R. K. Rainer, and T. E. Marshall. Business intelligence: An analysis of the literature. *Information Systems Management*, 25(2):121–131, 2008. doi: 10.1080/10580530801941512. URL http://dx.doi.org/10.1080/10580530801941512.

A. Jungherr, P. Jürgens, and H. Schoen. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t. o., sander, p. g., & welpe, i. m. "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review*, 30(2):229–234, 2012. doi: 10.1177/0894439311404119. URL http://ssc.sagepub.com/content/30/2/229.abstract.

L. Kagal, T. Finin, and A. Joshi. A policy based approach to security for the semantic web. *Proceedings of the 2nd International Semantic Web Conference*, 2870, 2003.

E. Kalampokis, M. Hausenblas, and K. Tarabanis. Combining social and government open data for participatory decision-making. In E. Tambouris, A. Macintosh, and H. Bruijn, editors, *Electronic Participation*, volume 6847 of *Lecture Notes in Computer Science*, pages 36–47. Springer Berlin Heidelberg, 2011a. ISBN 978-3-642-23332-6. doi: 10.1007/978-3-642-23333-3_4. URL http://dx.doi.org/10.1007/978-3-642-23333-3_4.

E. Kalampokis, E. Tambouris, and K. Tarabanis. A classification scheme for open government data: towards linking decentralised data. *Int. J. Web Eng. Technol.*, 6(3):266–285, June 2011b. ISSN 1476-1289. doi: 10.1504/IJWET.2011.040725. URL http://dx.doi.org/10.1504/IJWET.2011.040725.

E. Kalampokis, E. Tambouris, and K. Tarabanis. Open government data: A stage model. In M. Janssen, H. Scholl, M. Wimmer, and Y.-h. Tan, editors, *Electronic Government*, volume 6846 of *Lecture Notes in Computer Science*, pages 235–246. Springer Berlin Heidelberg, 2011c. ISBN 978-3-642-22877-3. doi: 10.1007/978-3-642-22878-0_20. URL http://dx.doi.org/10.1007/978-3-642-22878-0_20.

E. Kalampokis, A. Karamanou, E. Tambouris, and K. Tarabanis. Towards a vocabulary for incorporating predictive models into the linked data web. In CEUR-WS, editor, *First International Workshop on Semantic Statistics (SemStats) in conjuction with the in conjunction with the 12th International Semantic Web Conference (ISWC2013)*, 2013a.

E. Kalampokis, E. Tambouris, and K. Tarabanis. Understanding the predictive power of social media. *Internet Research*, 23(5):544–559, 2013b. doi: 10.1108/IntR-06-2012-0114. URL http://dx.doi.org/10.1108/IntR-06-2012-0114.

E. Kalampokis, E. Tambouris, and K. Tarabanis. Linked open government data analytics. In M. Wimmer, M. Janssen, and H. Scholl, editors, *Electronic Government*, volume 8074 of *Lecture Notes in Computer Science*, pages 99–110. Springer Berlin Heidelberg, 2013c. ISBN 978-3-642-40357-6. doi: 10.1007/978-3-642-40358-3_9. URL http://dx.doi.org/10.1007/978-3-642-40358-3_9.

E. Kalampokis, E. Tambouris, and K. Tarabanis. On publishing linked open government data. In *Proceedings of the 17th Panhellenic Conference on Informatics*, PCI '13, pages 25–32, New York, NY, USA, 2013d. ACM. ISBN 978-1-4503-1969-0. doi: 10.1145/2491845.2491869. URL http://doi.acm.org/10.1145/2491845.2491869.

E. Kalampokis, A. Nikolov, P. Haase, R. Cyganiak, A. Stasiewicz, A. Karamanou, M. Zotou, D. Zeginis, E. Tambouris, and K. Tarabanis. Exploiting linked data cubes with opencube toolkit. In M. Horridge, M. Rospocher, and J. van Ossenbruggen, editors, *Proc. of the ISWC 2014 Posters and Demos Track a track within 13th International Semantic Web Conference (ISWC2014)*, volume 1272. CEUR-WS, 2014. URL http://ceur-ws.org/Vol-1272/paper_109.pdf.

E. Kalampokis, B. Roberts, A. Karamanou, E. Tambouris, and K. Tarabanis. Challenges on developing tools for exploiting linked open data cubes. In *International Semantics Statistics Workshop 2015 (SemStats2015)*. CEUR-WS, 2015.

E. Kalampokis, A. Karamanou, E. Tambouris, and K. Tarabanis. What can twitter and linked open data reveal about elections results? the case of uk election, 2010. *Journal of Intelligent Information Systems*, Under Review, 2016a.

E. Kalampokis, E. Tambouris, A. Karamanou, and K. Tarabanis. Open statistics: The rise of a new era for open data? In *EGOV2016*. Springer, 2016b.

E. Kalampokis, E. Tambouris, and K. Tarabanis. Expanding data cubes for enhanced analytics on the web of linked data. *IEEE Transactions on Data and Knowledge Engineering*, under review, 2016c.

E. Kalampokis, E. Tambouris, and K. Tarabanis. Ict tools for creating, expanding, and exploiting statistical linked open data. *Statistical Journal of the IAOS*, accepted for publication, 2016d.

E. Kamateri, E. Kalampokis, E. Tambouris, and K. Tarabanis. The linked medical data access control framework. *Journal of Biomedical Informatics*, 50(0):213 – 225, 2014.

ISSN 1532-0464. doi: http://dx.doi.org/10.1016/j.jbi.2014.03.002. URL http://www.sciencedirect.com/science/article/pii/S1532046414000598. Special Issue on Informatics Methods in Medical Privacy.

B. Kämpgen. Dc proposal: Online analytical processing of statistical linked data. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist, editors, *The Semantic Web – ISWC 2011*, volume 7032 of *Lecture Notes in Computer Science*, pages 301–308. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-25092-7. doi: 10.1007/978-3-642-25093-4_22. URL http://dx.doi.org/10.1007/978-3-642-25093-4_22.

B. Kämpgen and A. Harth. No size fits all – running the star schema benchmark with sparql and rdf aggregate views. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 290–304. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-38287-1. doi: 10.1007/978-3-642-38288-8_20. URL http://dx.doi.org/10.1007/978-3-642-38288-8_20.

B. Kämpgen, S. O'Riain, and A. Harth. Interacting with statistical linked data via olap operations. In *C. Unger, P. Cimiano, editor, Proceedings of Interacting with Linked Data (ILD 2012), workshop co-located with the 9th Extended Semantic Web Conference*, pages 36–49. Citeseer, 2012.

B. Kämpgen, S. Stadtmüller, and A. Harth. Querying the global cube: Integration of multidimensional datasets from the web. In K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *Knowledge Engineering and Knowledge Management*, volume 8876 of *Lecture Notes in Computer Science*, pages 250–265. Springer International Publishing, 2014. ISBN 978-3-319-13703-2. doi: 10.1007/978-3-319-13704-9_20. URL http://dx.doi.org/10.1007/978-3-319-13704-9_20.

M. Kaufmann and J. Kalita. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*, 2010.

D. B. Keator, D. Wei, S. Gadde, J. Bockholt, J. S. Grethe, D. Marcus, N. Aucoin, and I. B. Ozyurt. Derived Data Storage and Exchange Workflow for Large-Scale Neuroimaging Analyses on the BIRN Grid. *Frontiers in neuroinformatics*, 3, 2009.

ISSN 1662-5196. doi: 10.3389/neuro.11.030.2009. URL http://dx.doi.org/10.3389/neuro.11.030.2009.

T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. A. Evelo, and A. R. Pico. Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(Database-Issue):1301–1307, 2012. doi: 10.1093/nar/gkr1074. URL http://dx.doi.org/10.1093/nar/gkr1074.

M. Khalil, B. Lanvin, and V. Chaudhry. The e-government handbook for developing countries: A project of infodev and the center for democracy and technology. Technical report, Washington, DC: InfoDev and the Center for Democracy & Technology, November 2002.

R. Kimball and M. Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

S. Kinsella, M. Wang, J. Breslin, and C. Hayes. Improving categorisation in social media using hyperlinks to structured data sources. In G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, editors, *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 390–404. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-21063-1. doi: 10.1007/978-3-642-21064-8_27. URL http://dx.doi.org/10.1007/978-3-642-21064-8_27.

S. Kirrane, A. Abdelrahman, A. Mileo, and S. Decker. Secure manipulation of linked data. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 248–263, 2013. doi: 10.1007/978-3-642-41335-3_16. URL http://dx.doi.org/10.1007/978-3-642-41335-3_16.

I. S. Kohane, S. E. Churchill, and S. N. Murphy. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc*, 19:181–5, 2012 Mar-Apr 2012. ISSN 1527-974X. doi: 10.1136/amiajnl-2011-000492.

S. Konishi and G. Kitagawa. *Information criteria and statistical modeling*. Springer, 2008.

J. Krauss, S. Nann, and D. Simon. Predicting movie success and academy awards through sentiment and social network analysis. In *16th European Conference on Information Systems*, pages 2026–2037, 2008.

S. Kulk and B. Van Loenen. Brave new open data world? *International Journal of Spatial Data Infrastructures Research*, 7:196–206, 2012.

R. D. Kush, E. Helton, F. W. Rockhold, and C. D. Hardison. Electronic health records, medical research, and the tower of babel. *New England Journal of Medicine*, 358(16): 1738–1740, 2008.

H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772751. URL http://doi.acm.org/10.1145/1772690.1772751.

V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.*, 3(4):72:1–72:22, Sept. 2012. ISSN 2157-6904. doi: 10.1145/2337542.2337557. URL http://doi.acm.org/10.1145/2337542.2337557.

E. C. Lau, F. S. Mowat, M. A. Kelsh, J. C. Legg, N. M. Engel-Nitz, H. N. Watson, H. L. Collins, R. J. Nordyke, and J. L. Whyte. Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clinical Epidemiology*, 3:259–272, 2011. ISSN 1179-1349. doi: 10.2147/CLEP.S23690. URL http://dx.doi.org/10.2147/CLEP.S23690.

K. Layne and J. Lee. Developing fully functional e-government: A four stage model. *Government Information Quarterly*, 18(2):122 – 136, 2001. ISSN 0740-624X. doi: http://dx.doi.org/10.1016/S0740-624X(01)00066-1. URL http://www.sciencedirect.com/science/article/pii/S0740624X01000661.

J. Lee. 10 year retrospect on stage models of e-government: A qualitative meta-synthesis. *Government Information Quarterly*, 27(3):220 – 230, 2010. ISSN 0740-624X. doi: http://dx.doi.org/10.1016/j.giq.2009.12.009. URL http://www.sciencedirect.com/science/article/pii/S0740624X10000249.

M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA, 2002. ACM. ISBN 1-58113-507-6. doi: 10.1145/543613.543644. URL http://doi.acm.org/10.1145/543613.543644.

D. Levy, G. B. Ehret, K. Rice, G. C. Verwoert, L. J. Launer, A. Dehghan, N. L. Glazer, A. C. Morrison, A. D. Johnson, T. Aspelund, Y. Aulchenko, T. Lumley, A. Köttgen, R. S. Vasan, F. Rivadeneira, G. Eiriksdottir, X. Guo, D. E. Arking, G. F. Mitchell, F. U. S. Mattace-Raso, A. V. Smith, K. Taylor, R. B. Scharpf, S.-J. Hwang, E. J. G. Sijbrands, J. Bis, T. B. Harris, S. K. Ganesh, C. J. O'Donnell, A. Hofman, J. I. Rotter, J. Coresh, E. J. Benjamin, A. G. Uitterlinden, G. Heiss, C. S. Fox, J. C. M. Witteman, E. Boerwinkle, T. J. Wang, V. Gudnason, M. G. Larson, A. Chakravarti, B. M. Psaty, and C. M. van Duijn. Genome-wide association study of blood pressure and hypertension. *Nat Genet*, 41:677–87, 2009 Jun 2009. ISSN 1546-1718. doi: 10.1038/ng.384.

X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002519.

Y. Liu, X. Huang, A. An, and X. Yu. Arsa: A sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 607–614, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277845. URL http://doi.acm.org/10.1145/1277741.1277845.

Y. Liu, Y. Chen, R. Lusch, H. Chen, D. Zimbra, and S. Zeng. User-generated content on social media: Predicting market success with online word-of-mouth. *IEEE Intell. Syst*, 25(1):75–78, 2010.

A. Livne, M. P. Simmons, E. Adar, and L. A. Adamic. The party is over here: Structure and content in the 2010 election. In *ICWSM*, 2011.

B. Lorincz, G. Colclough, D. Tinholt, C. van Oranje, G. Cattaneo, and L. Jacquet. Smarter, faster, better egovernment: 8th benchmark measurement. Technical report, European Commission, Directorate General for Information Society and Media, November 2009.

C. S. Louis and G. Zorlu. Can twitter predict disease outbreaks? *BMJ*, 344, 5 2012. doi: 10.1136/bmj.e2353.

E. J. Ludman, S. M. Fullerton, L. Spangler, S. B. Trinidad, M. M. Fujii, G. P. Jarvik, E. B. Larson, and W. Burke. Glad you asked: participants' opinions of re-consent for dbgap data submission. *Journal of empirical research on human research ethics: JERHRE*, 5(3):9, 2010.

C. Lui, P. T. Metaxas, and E. Mustafaraj. On the predictability of the us elections through search volume activity. In *IADIS International Conference e-Society*, pages 165–172, 2011.

C. Mader, M. Martin, and C. Stadler. Facilitating the exploration and visualization of linked data. In S. Auer, V. Bryl, and S. Tramp, editors, *Linked Open Data – Creating Knowledge Out of Interlinked Data*, volume 8661 of *Lecture Notes in Computer Science*, pages 90–107. Springer International Publishing, 2014. ISBN 978-3-319-09845-6. doi: 10.1007/978-3-319-09846-3_5. URL http://dx.doi.org/10.1007/978-3-319-09846-3_5.

B. Malin, D. Karp, and R. H. Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of investigative medicine: the official publication of the American Federation for Clinical Research*, 58(1):11, 2010.

J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

J. Manyika, M. Chui, P. Groves, D. Farrell, S. V. Kuiken, and E. A. Doshi. Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey & Company, October 2013. URL http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.

J. C. Maro, R. Platt, J. H. Holmes, B. L. Strom, S. Hennessy, R. Lazarus, and J. S. Brown. Design of a national distributed health data network. *Annals of Internal Medicine*, 151(5):341–344, 2009. doi: 10.7326/0003-4819-151-5-200909010-00139. URL http://dx.doi.org/10.7326/0003-4819-151-5-200909010-00139.

S. Martin, M. Foulonneau, S. Turki, and M. Ihadjadene. Open data: barriers, risks and opportunities. In *Proceedings of the 13th European Conference on eGovernment: ECEG*, pages 301–309, 2013.

E. Marx, S. Shekarpour, S. Auer, and A.-C. Ngomo. Large-scale rdf dataset slicing. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pages 228–235, Sept 2013. doi: 10.1109/ICSC.2013.47.

A. J. McMurry, S. N. Murphy, D. MacFadden, G. Weber, W. W. Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, S. Churchill, and I. S. Kohane. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. *PLoS ONE*, 8(3): e55811+, Mar. 2013. doi: 10.1371/journal.pone.0055811. URL http://dx.doi.org/10.1371/journal.pone.0055811.

A. Meijer, J. de Hoog, M. van Twist, M. van der Steen, and J. Scherpenisse. Understanding the dynamics of open data: From sweeping statements to complex contextual interactions. In M. Gascó-Hernández, editor, *Open Government*, volume 4 of *Public Administration and Information Technology*, pages 101–114. Springer New York, 2014. ISBN 978-1-4614-9562-8. doi: 10.1007/978-1-4614-9563-5_7. URL http://dx.doi.org/10.1007/978-1-4614-9563-5_7.

P. Mendes, A. Passant, P. Kapanipathi, and A. Sheth. Linked open social signals. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 224–231, Aug 2010. doi: 10.1109/WI-IAT.2010.314.

A. Meroño-Peñuela, A. Ashkpour, L. Rietveld, and R. Hoekstra. Linked humanities data: The next frontier? a case-study in historical census data. In *Proceedings of the 2nd International Workshop on Linked Science 2012*, volume 951, 2012.

P. Metaxas, E. Mustafaraj, and D. Gayo-Avello. How (not) to predict elections. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on*

and 2011 ieee third international conference on social computing (socialcom), pages 165 –171, oct. 2011. doi: 10.1109/PASSAT/SocialCom.2011.98.

A. Miles and S. Bechhofer. SKOS simple knowledge organization system. Technical report, W3C, August 2009.

G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 155–158, 2006.

J. C. Molloy. The open knowledge foundation: Open data means better science. *PLoS Biol*, 9(12):e1001195, 12 2011. doi: 10.1371/journal.pbio.1001195. URL http://dx.doi.org/10.1371%2Fjournal.pbio.1001195.

M. Morales-Arroyo and T. Pandey. Identification of critical ewom dimensions for music albums. In *Management of Innovation and Technology (ICMIT), 2010 IEEE International Conference on*, pages 1230 –1235, june 2010. doi: 10.1109/ICMIT.2010.5492860.

S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association : JAMIA*, 17(2):124–130, Mar. 2010. ISSN 1527-974X. doi: 10.1136/jamia.2009.000893. URL http://dx.doi.org/10.1136/jamia.2009.000893.

P. Murray-Rust. Open data in science. *Serials Review*, 34(1):52–64, 2008. doi: 10.1080/00987913.2008.10765152. URL http://www.tandfonline.com/doi/abs/10.1080/00987913.2008.10765152.

M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 189–192, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718953. URL http://doi.acm.org/10.1145/1718918.1718953.

D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007. doi: doi:10.1075/li.30.1.03nad. URL http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002.

V. Nebot and R. Berlanga. Building data warehouses with semantic web data. *Decision Support Systems*, 52(4):853 – 868, 2012. ISSN 0167-9236. doi: http://dx.doi.org/10.1016/j.dss.2011.11.009. URL http://www.sciencedirect.com/science/article/pii/S0167923611002156. 1)Decision Support Systems for Logistics and Supply Chain Management 2)Business Intelligence and the Web.

C. Newton-Cheh, T. Johnson, V. Gateva, M. D. D. Tobin, M. Bochud, L. Coin, S. S. S. Najjar, Jing, S. C. C. Heath, S. Eyheramendy, K. Papadakis, B. F. F. Voight, L. J. J. Scott, F. Zhang, M. Farrall, T. Tanaka, C. Wallace, J. C. C. Chambers, K.-T. T. Khaw, P. Nilsson, P. van der Harst, S. Polidoro, D. E. E. Grobbee, Charlotte, M. L. L. Bots, L. V. V. Wain, K. S. S. Elliott, A. Teumer, J. Luan, G. Lucas, J. Kuusisto, P. R. R. Burton, D. Hadley, W. L. L. Mcardle, M. Brown, A. Dominiczak, S. J. J. Newhouse, N. J. J. Samani, J. Webster, E. Zeggini, J. S. S. Beckmann, S. Bergmann, N. Lim, K. Song, P. Vollenweider, G. Waeber, D. M. M. Waterworth, X. Yuan, L. Groop, M. Orho-Melander, A. Allione, A. Di Gregorio, S. Guarrera, S. Panico, F. Ricceri, V. Romanazzi, C. Sacerdote, P. Vineis, I. Barroso, M. S. S. Sandhu, R. N. N. Luben, G. J. J. Crawford, P. Jousilahti, M. Perola, M. Boehnke, L. L. L. Bonnycastle, F. S. S. Collins, A. U. U. Jackson, K. L. L. Mohlke, H. M. M. Stringham, T. T. T. Valle, C. J. J. Willer, R. N. N. Bergman, M. A. A. Morken, A. Döring, C. Gieger, T. Illig, T. Meitinger, E. Org, A. Pfeufer, Erich, S. Kathiresan, J. Marrugat, C. J. J. O'Donnell, S. M. M. Schwartz, D. S. S. Siscovick, I. Subirana, N. B. B. Freimer, A.-L. L. Hartikainen, M. I. I. Mccarthy, P. F. F. O'Reilly, L. Peltonen, A. Pouta, P. E. E. de Jong, H. Snieder, W. H. H. van Gilst, R. Clarke, A. Goel, A. Hamsten, J. F. F. Peden, U. Seedorf, A.-C. C. Syvänen, G. Tognoni, E. G. G. Lakatta, S. Sanna, P. Scheet, D. Schlessinger, A. Scuteri, M. Dörr, F. Ernst, S. B. B. Felix, G. Homuth, R. Lorbeer, T. Reffelmann, R. Rettig, U. Völker, P. Galan, I. G. G. Gut, S. Hercberg, Mark, D. Zelenika, P. Deloukas, N. Soranzo, F. M. M. Williams, G. Zhai, V. Salomaa, M. Laakso, R. Elosua, N. G. G. Forouhi, H. Völzke, C. S. S. Uiterwaal, Y. T. T. van der Schouw, M. E. E. Numans, G. Matullo, G. Navis, G. Berglund, S. A. A. Bingham, J. S. S. Kooner, J. M. M. Connell, S. Bandinelli, L. Ferrucci, H. Watkins, T. D. D. Spector, J. Tuomilehto, D. Altshuler, D. P. P. Strachan, M. Laan, P. Meneton, N. J. J. Wareham, M. Uda, M.-R. R. Jarvelin, V. Mooser, O. Melander, Ruth, P. Elliott, G. R. R. Abecasis, M. Caulfield, and P. B. B. Munroe. Genome-wide association study identifies eight loci associated with blood pressure. *Nature genetics*, 41(6):666–676, June 2009.

ISSN 1546-1718. doi: 10.1038/ng.361. URL http://dx.doi.org/10.1038/ng.361.

T. Niemi, L. Hirvonen, and K. Järvelin. Multidimensional data model and query language for informetrics. *Journal of the American Society for Information Science and Technology*, 54(10):939–951, 2003.

B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

O. of Fair Trading. The commercial use of public informaiton (cupi), December 2006.

A. Oghina, M. Breuss, M. Tsagkias, and M. Rijke. Predicting imdb movie ratings using social media. In R. Baeza-Yates, A. Vries, H. Zaragoza, B. Cambazoglu, V. Murdock, R. Lempel, and F. Silvestri, editors, *Advances in Information Retrieval*, volume 7224 of *Lecture Notes in Computer Science*, pages 503–507. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28996-5. doi: 10.1007/978-3-642-28997-2_51. URL http://dx.doi.org/10.1007/978-3-642-28997-2_51.

C. Oh and O. Sheng. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In *ICIS*, number 17, 2011.

L. Ohno-Machado, V. Bafna, A. A. Boxwala, B. E. Chapman, W. W. Chapman, K. Chaudhuri, M. E. Day, C. Farcas, N. D. Heintzman, X. Jiang, H. Kim, J. Kim, M. E. Matheny, F. S. Resnic, and S. A. Vinterbo. idash: integrating data for analysis, anonymization, and sharing. *J Am Med Inform Assoc*, 2011 Nov 10 2011. ISSN 1527-974X. doi: 10.1136/amiajnl-2011-000538.

T. B. Pedersen, C. S. Jensen, and C. E. Dyreson. A foundation for capturing and querying complex multidimensional data. *Information Systems*, 26(5):383 – 423, 2001. ISSN 0306-4379. doi: 10.1016/S0306-4379(01)00023-0. URL http://www.sciencedirect.com/science/article/pii/S0306437901000230. Data Warehousing.

H. E. Pence and A. Williams. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.*, 87(11):1123–1124, Aug. 2010. doi: 10.1021/ed100697w. URL http://dx.doi.org/10.1021/ed100697w.

K. Perakis, T. Bouras, D. Ntalaperas, P. Hasapis, C. Georgousopoulos, R. Sahay, O. D. Beyan, C. Potlog, and D. Usurelu. Advancing patient record safety and ehr semantic interoperability. In *SMC'13*, pages 3251–3257, 2013.

V. Peristeras and K. Tarabanis. Towards an enterprise architecture for public administration using a top-down approach. *European Journal of Information Systems*, 9(4):252–260, 2000. URL http://www.ingentaconnect.com/content/pal/0960085x/2000/00000009/00000004/3000378.

I. Petrou, G. Papastefanatos, and T. Dalamagas. Publishing census as linked open data: A case study. In *Proceedings of the 2Nd International Workshop on Open Data*, WOD '13, pages 4:1–4:3, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2020-7. doi: 10.1145/2500410.2500412. URL http://doi.acm.org/10.1145/2500410.2500412.

P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, 2008. doi: 10.1086/593098. URL http://cid.oxfordjournals.org/content/47/11/1443.abstract.

H. U. Prokosch and T. Ganslandt. Perspectives for Medical Informatics. Reusing the Electronic Medical Record for Clinical Research. *Methods of Information in Medicine*, 48:38–44, 2009.

J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, 2011.

F. Ravat, O. Teste, R. Tournier, and G. Zurfluh. Algebraic and graphic languages for olap manipulations. *International Journal of Data Warehousing and Mining (IJDWM)*, 4 (1):17–46, 2008.

D. A. Reinstein and C. M. Snyder. The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The Journal of Industrial Economics*, 53(1):27–51, 2005. ISSN 1467-6451. doi: 10.1111/j.0022-1821.2005.00244.x. URL http://dx.doi.org/10.1111/j.0022-1821.2005.00244.x.

J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *Proceedings of the 1st International Workshop on Mining Social Media*, pages 9–17, 2009.

D. Robinson, H. Yu, W. P. Zeller, and E. W. Felten. Government data and the invisible hand. *Yale JL & Tech.*, 11:159, 2008.

O. Romero and A. Abelló. A survey of multidimensional modeling methodologies. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(2):1–23, 2009.

M. Rowe and M. Stankovic. Aligning tweets with events: Automation via semantics. *Semantic Web*, 3(2):115–130, 2012.

L. Ruback, M. Pesce, S. Manso, S. Ortiga, P. E. R. Salas, and M. A. Casanova. A mediator for statistical linked data. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 339–341, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1656-9. doi: 10.1145/2480362.2480432. URL http://doi.acm.org/10.1145/2480362.2480432.

H. Rui and A. Whinston. Designing a social-broadcasting-based business intelligence system. *ACM Trans. Manage. Inf. Syst.*, 2(4):22:1–22:19, Jan. 2012. ISSN 2158-656X. doi: 10.1145/2070710.2070713. URL http://doi.acm.org/10.1145/2070710.2070713.

O. Sacco, A. Passant, and S. Decker. An access control framework for the web of data. *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 0:456–463, 2011. doi: http://doi.ieeecomputersociety.org/10.1109/TrustCom.2011.59.

S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. Thibodeau Jr, S. Auer, J. Sequeda, and A. Ezzat. A survey of current approaches for mapping of relational databases to rdf. Technical report, W3C RDB2RDF Incubator Group, 2009.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772777. URL http://doi.acm.org/10.1145/1772690.1772777.

P. Salas, M. Martin, F. Mota, S. Auer, K. Breitman, and M. Casanova. Olap2datacube: An ontowiki plug-in for statistical data publishing. In *Developing Tools as Plug-ins (TOPI), 2012 2nd Workshop on*, pages 79–83, June 2012a. doi: 10.1109/TOPI.2012.6229815.

P. E. R. Salas, F. M. Da Mota, K. K. Breitman, M. A. Casanova, M. Martin, and S. Auer. Publishing statistical data on the web. *International Journal of Semantic Computing*, 06(04):373–388, 2012b. doi: 10.1142/S1793351X12400119. URL http://www.worldscientific.com/doi/abs/10.1142/S1793351X12400119.

I. Sanderson. Evaluation, policy learning and evidence-based policy making. *Public Administration*, 80(1):1–22, 2002. ISSN 1467-9299. doi: 10.1111/1467-9299.00292. URL http://dx.doi.org/10.1111/1467-9299.00292.

E. T. K. Sang and J. Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2389969.2389976.

S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C. T. Evelo, M. J. Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S. J. H. Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide. Toward interoperable bioscience data. *Nat Genet*, 44(2):121–126, 02 2012. URL http://dx.doi.org/10.1038/ng.1054.

M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, editors, *The Semantic Web – ISWC 2014*, volume 8796 of *Lecture Notes in Computer Science*, pages 245–260. Springer International Publishing, 2014. ISBN 978-3-319-11963-2. doi: 10.1007/978-3-319-11964-9_16. URL http://dx.doi.org/10.1007/978-3-319-11964-9_16.

P. N. Schofield, T. Bubela, T. Weaver, L. Portilla, S. D. Brown, J. M. Hancock, D. Einhorn, G. Tocchini-Valentini, M. Hrabe de Angelis, and N. Rosenthal. Post-publication sharing of data and tools. *Nature*, 461(7261):171–173, 09 2009. URL http://dx.doi.org/10.1038/461171a.

A. L. Sherborne, K. Hemminki, R. Kumar, C. R. Bartram, M. Stanulla, M. Schrappe, E. Petridou, Á. F. Semsei, C. Szalai, D. Sinnett, M. Krajinovic, J. Healy, M. Lanciotti, C. Dufour, S. Indaco, E. A. El-Ghouroury, R. Sawangpanich, S. Hongeng, S. Pakakasama, A. Gonzalez-Neira, E. L. Ugarte, V. P. Leal, J. P. Espinoza, A. M. Kamel, G. T. Ebid, E. R. Radwan, S. Yalin, E. Yalin, M. Berkoz, J. Simpson, E. Roman, T. Lightfoot, F. J. Hosking, J. Vijayakrishnan, M. Greaves, and R. S. Houlston. Rationale for an international consortium to study inherited genetic susceptibility to childhood acute lymphoblastic leukemia. *Haematologica*, 96(7):1049–1054, 2011. ISSN 0390-6078. doi: 10.3324/haematol.2011.040121.

P. Shi. Guiding school choice reform through novel applications of operations research. Technical report, MIT Operations Research Center, September 2013. URL http://www.mit.edu/~pengshi/papers/guiding-reform.pdf.

G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 08 2010. doi: 10.1214/10-STS330. URL http://dx.doi.org/10.1214/10-STS330.

G. Shmueli and O. R. Koppius. Predictive analytics in information systems research. *MIS Quarterly*, 35(3):553 – 572, 2011. ISSN 02767783. URL http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=63604908&site=ehost-live.

K. Siau and Y. Long. Synthesizing e-government stage models – a meta-synthesis based on meta-ethnography approach. *Industrial Management & Data Systems*, 105(4): 443–458, 2005.

A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5):e19467, 05 2011. doi: 10.1371/journal.pone.0019467. URL http://dx.doi.org/10.1371%2Fjournal.pone.0019467.

M. Skoric, N. Poor, P. Achananuparp, E.-P. Lim, and J. Jiang. Tweets and votes: A study of the 2011 singapore general election. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2583–2591, Jan 2012. doi: 10.1109/HICSS.2012.607.

M. Slaymaker, D. Power, D. Russell, G. Wilson, and A. Simpson. Accessing and aggregating legacy data sources for healthcare research, delivery and training. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages

1317–1324, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-753-7. doi: 10.1145/1363686.1363994. URL http://doi.acm.org/10.1145/1363686.1363994.

D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Commun. ACM*, 40(5):103–110, May 1997. ISSN 0001-0782. doi: 10.1145/253769.253804. URL http://doi.acm.org/10.1145/253769.253804.

E. Tambouris. An integrated platform for realising online one-stop government: the egov project. In *Database and Expert Systems Applications, 2001. Proceedings. 12th International Workshop on*, pages 359–363, 2001. doi: 10.1109/DEXA.2001.953087.

E. Tambouris, E. Kalampokis, and K. Tarabanis. Processing linked open data cubes. In E. Tambouris, M. Janssen, H. J. Scholl, M. A. Wimmer, K. Tarabanis, M. Gascó, B. Klievink, I. Lindgren, and P. Parycek, editors, *Electronic Government*, volume 9248 of *Lecture Notes in Computer Science*, pages 130–143. Springer International Publishing, 2015. ISBN 978-3-319-22478-7. doi: 10.1007/978-3-319-22479-4_10. URL http://dx.doi.org/10.1007/978-3-319-22479-4_10.

P. Taylor. Personal Genomes: When consent gets in the way. *Nature*, 456(7218):32–33, Nov. 2008. ISSN 0028-0836. doi: 10.1038/456032a. URL http://dx.doi.org/10.1038/456032a.

T. Tong and H. Zhao. Practical guidelines for assessing power and false discovery rate for a fixed sample size in microarray experiments. *Stat Med*, Mar. 2008. ISSN 0277-6715. doi: 10.1002/sim.3237. URL http://dx.doi.org/10.1002/sim.3237.

M. Tsagkias, W. Weerkamp, and M. Rijke. News comments:exploring, modeling, and online prediction. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. Rijsbergen, editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*, pages 191–203. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-12274-3. doi: 10.1007/978-3-642-12275-0_19. URL http://dx.doi.org/10.1007/978-3-642-12275-0_19.

F. S. Tseng and C.-W. Chen. Integrating heterogeneous data warehouses using xml technologies. *Journal of Information Science*, 31(3):209–229, 2005. doi: 10.1177/0165551505052467. URL http://jis.sagepub.com/content/31/3/209.abstract.

T. Tudorache, C. Nyulas, N. F. Noy, and M. A. Musen. Using semantic web in ICD-11: three years down the road. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*, pages 195–211, 2013. doi: 10.1007/978-3-642-41338-4_13. URL http://dx.doi.org/10.1007/978-3-642-41338-4_13.

A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 178–185, 2010.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Where there is a sea there are pirates: Response to jungherr, jürgens, and schoen. *Social Science Computer Review*, 30(2):235–239, 2012. doi: 10.1177/0894439311404123. URL http://ssc.sagepub.com/content/30/2/235.abstract.

G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. Sig.ma: Live views on the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):355 – 364, 2010. ISSN 1570-8268. doi: http://dx.doi.org/10.1016/j.websem.2010.08.003. URL http://www.sciencedirect.com/science/article/pii/S1570826810000624. Semantic Web Challenge 2009 User Interaction in Semantic Web research.

A. F. van Veenstra and T. van den Broek. A community-driven open data lifecycle model based on literature and practice. In I. Boughzala, M. Janssen, and S. Assar, editors, *Case Studies in e-Government 2.0*, pages 183–198. Springer International Publishing, 2015. ISBN 978-3-319-08080-2. doi: 10.1007/978-3-319-08081-9_11. URL http://dx.doi.org/10.1007/978-3-319-08081-9_11.

P. Vassiliadis. Modeling multidimensional databases, cubes and cube operations. In *Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on*, pages 53–62, Jul 1998. doi: 10.1109/SSDM.1998.688111.

G. Vickery. Review of recent studies on psi re-use and related market developments. *Information Economics, Paris*, 2011.

B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. Methodological guidelines for publishing government linked data. In D. Wood, editor, *Linking Government Data*, pages 27–49. Springer New York, 2011. ISBN 978-1-4614-1766-8. doi: 10.1007/978-1-4614-1767-5_2. URL http://dx.doi.org/10.1007/978-1-4614-1767-5_2.

J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Silk-a link discovery framework for the web of data. In *Proceedings of the 2nd Linked Data on the Web Workshop (LDOW2009)*, 2009.

S. Vosen and T. Schmidt. Forecasting private consumption: survey-based indicators vs. google trends. *Journal of Forecasting*, 30(6):565–578, 2011. ISSN 1099-131X. doi: 10.1002/for.1213. URL http://dx.doi.org/10.1002/for.1213.

S. Vosen and T. Schmidt. A monthly consumption indicator for germany based on internet search query data. *Applied Economics Letters*, 19(7):683–687, 2012. doi: 10.1080/13504851.2011.595673. URL http://www.tandfonline.com/doi/abs/10.1080/13504851.2011.595673.

A. H. Wang. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, pages 1–10, July 2010.

X. Wang, M. Gerber, and D. Brown. Automatic crime prediction using events extracted from twitter posts. In S. Yang, A. Greenberg, and M. Endsley, editors, *Social Computing, Behavioral - Cultural Modeling and Prediction*, volume 7227 of *Lecture Notes in Computer Science*, pages 231–238. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-29046-6. doi: 10.1007/978-3-642-29047-3_28. URL http://dx.doi.org/10.1007/978-3-642-29047-3_28.

G. M. Weber, S. N. Murphy, A. J. McMurry, D. Macfadden, D. J. Nigrin, S. Churchill, and I. S. Kohane. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. *Journal of the American Medical Informatics Association : JAMIA*, 16(5):624–630, June 2009. ISSN 1527-974X. doi: 10.1197/jamia.m3191. URL http://dx.doi.org/10.1197/jamia.m3191.

J. Webster and R. T. Watson. Analyzing the past to prepare for the future: writing a literature review. *MIS Q.*, 26(2):xiii–xxiii, June 2002. ISSN 0276-7783. URL http://dl.acm.org/citation.cfm?id=2017160.2017162.

M. G. Weiner and P. J. Embi. Toward Reuse of Clinical Data for Research and Quality Improvement: The End of the Beginning? *Annals of Internal Medicine*, 151(5): 359–360, Sept. 2009. doi: 10.1059/0003-4819-151-5-200909010-00141. URL http://dx.doi.org/10.1059/0003-4819-151-5-200909010-00141.

D. M. West. E-government and the transformation of service delivery and citizen attitudes. *Public Administration Review*, 64(1):15–27, 2004. ISSN 1540-6210. doi: 10.1111/j.1540-6210.2004.00343.x. URL http://dx.doi.org/10.1111/j.1540-6210.2004.00343.x.

E. L. Willighagen, A. Waagmeester, O. Spjuth, P. Ansell, A. J. Williams, V. Tkachenko, J. Hastings, B. Chen, and D. J. Wild. The chembl database as linked open data. *J. Cheminformatics*, 5:23, 2013. doi: 10.1186/1758-2946-5-23. URL http://dx.doi.org/10.1186/1758-2946-5-23.

K. Wilson and J. S. Brownstein. Early detection of disease outbreaks using the internet. *Canadian Medical Association Journal*, 180(8):829–831, 2009. doi: 10.1503/cmaj.1090215. URL http://www.cmaj.ca/content/180/8/829.short.

L. Wu and E. Brynjolfsson. The future of prediction: how google searches foreshadow housing prices and quantities. In *30th International Conference on Information Systems*. AISE, 2009.

S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963504. URL http://doi.acm.org/10.1145/1963405.1963504.

Q. Ye, R. Law, and B. Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180 – 182, 2009. ISSN 0278-4319. doi: 10.1016/j.ijhm.2008.06.011. URL http://www.sciencedirect.com/science/article/pii/S0278431908000546.

B. Zapilko and B. Mathiak. Performing statistical methods on linked data. *International Conference on Dublin Core and Metadata Applications*, 0, 2011. ISSN 1939-1366. URL http://dcpapers.dublincore.org/pubs/article/view/3627.

A. Zaveri, J. Nickenig Vissoci, C. Daraio, and R. Pietrobon. Using linked data to evaluate the impact of research and development in europe: A structural equation model. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. Parreira, L. Aroyo, N. Noy, C. Welty, and K. Janowicz, editors, *The Semantic Web – ISWC 2013*, volume 8219 of *Lecture Notes in Computer Science*, pages 244–259. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-41337-7. doi: 10.1007/978-3-642-41338-4_16. URL http://dx.doi.org/10.1007/978-3-642-41338-4_16.

X. Zhang, H. Fuehres, and P. A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26(0):55 – 62, 2011. ISSN 1877-0428. doi: http://dx.doi.org/10.1016/j.sbspro.2011.10.562. URL http://www.sciencedirect.com/science/article/pii/S1877042811023895. The 2nd Collaborative Innovation Networks Conference - {COINs2010}.

X. Zhang, H. Fuehres, and P. Gloor. Predicting asset value through twitter buzz. In J. Altmann, U. Baumöl, and B. J. Krämer, editors, *Advances in Collective Intelligence 2011*, volume 113 of *Advances in Intelligent and Soft Computing*, pages 23–34. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-25320-1. doi: 10.1007/978-3-642-25321-8_3. URL http://dx.doi.org/10.1007/978-3-642-25321-8_3.

P. Zhao, X. Li, D. Xin, and J. Han. Graph cube: On warehousing and olap multidimensional networks. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 853–864, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0661-4. doi: 10.1145/1989323.1989413. URL http://doi.acm.org/10.1145/1989323.1989413.

A. Zuiderwijk and M. Janssen. Barriers and development directions for the publication and usage of open data: A socio-technical view. In M. Gascó-Hernández, editor, *Open Government*, volume 4 of *Public Administration and Information Technology*, pages 115–135. Springer New York, 2014. ISBN 978-1-4614-9562-8. doi: 10.1007/978-1-4614-9563-5_8. URL http://dx.doi.org/10.1007/978-1-4614-9563-5_8.