Profiling User Interests on the Social Semantic Web

Fabrizio Orlandi

Submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

SUPERVISOR:
Dr. Alexandre Passant
Dr. John G. Breslin

INTERNAL EXAMINER:
Prof. Dr. Stefan Decker

EXTERNAL EXAMINER:
Dr. Fabien Gandon

Digital Enterprise Research Institute,
National University of Ireland, Galway
March 2014

Abstract

The World Wide Web is evolving towards an ecosystem of applications and services offering personalised content to its users. At the same time, the widespread adoption of social media led its users to provide portions of their personal data on several different services for socialisation or personalisation purposes. The automated extraction of users' interests from personal Social Web data is becoming an essential part of the current Web applications for personalisation and recommendation. Such personalisation is required in order to provide an adaptive Web to users, where content fits their preferences, background and current interests, making the Web more social and relevant. Current techniques of personalisation systems analyse user activities on a social media system and collect sets of tags, entities or links to represent users' interests. These sets representing users' interests, also called *user profiles of interests*, are often missing a deeper "understanding" of the represented interests. Moreover, these user profiles cannot be easily exchanged between social media systems, therefore lacking portability and interoperability of personal user information. As a remedy, we propose a complete methodology for profiling user interests that leverages Semantic Web technologies and provenance of Social Web data.

The Semantic Web represents a prominent recent approach attempting to provide the Web with a meaning not only people, but also machines can process. We adopt Semantic Web technologies for creating a standard interoperable representation of user profiles of interests. This allows for aggregation of heterogeneous user models from different social websites, and knowledge enrichment about user entities of interest. Moreover, we leverage provenance management of Social Web data to retrieve complete information about data producers (either applications, software agents or users) and increase the accuracy of user profiles. Provenance of data can be considered as one of the core building blocks for establishing data quality measures, for enhancing the knowledge acquisition/filtering process, and the user profiling phase. We investigate and evaluate a set of heuristics for mining users' interests from their social activities on heterogeneous social media websites and propose different approaches and measures for aggregating, enriching and ranking users' concepts of interest. Finally, we evaluate our methodology for profiling user interests in a practical Web personalisation scenario.

Declaration

I declare that this thesis is composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. The work reported in this thesis was supported by Science Foundation Ireland (SFI) under the DERI-Líon II project (SFI/08/CE/I1380), and by an IRCSET scholarship.

Fabrizio Orlandi

Acknowledgements

There are many people I would like to thank for their support over the past four years. First, I would like to thank Alex who gave me the opportunity to come to Galway and DERI: he believed in me since my first internship and guided me through many obstacles till the end of the Ph.D. In the same way, I would like to thank John for his advice and encouragement especially over the last two years. Big thanks also to my examiners Stefan and Fabien for their valuable feedback and discussions. I am grateful to my colleagues at DERI and Kno.e.sis for the many inspiring conversations and exciting activities. Special thanks to those who also shared moments of fun and frustration with me at work: some of them became very good friends and, together with other friends from all over the world, contributed in making these last four years a memorable experience. Finally, last but not least, I would like to mention those who simply gave me everything they could to support me and were always there for me: Lisa, my sister and especially my parents, who sacrificed much of their life for me.

Grazie!
Thank you!
Go raibh maith agaibh!
Danke!

Publications

The work described in this thesis was partially covered by or stemmed from the following publications and presentations:

- *Journal or magazine articles*:

  1. Owen Sacco, **Fabrizio Orlandi**, Alexandre Passant. *Privacy aware and faceted user profile management using social data.* In Semantic Web Journal, 2012. Accepted with revision (available online from 08 April 2011).

  2. **Fabrizio Orlandi**, Alexandre Passant. *Modelling provenance of DBpedia resources using Wikipedia contributions.* In Web Semantics: Science, Services and Agents on the World Wide Web, 9(2), 149 — 164, Elsevier, 2011.

  3. **Fabrizio Orlandi**, Pavan Kapanipathi, Amit Sheth, Alexandre Passant. *Real-time semantic personalisation of Social Web streams.* Journal submission under review, 2014.

- *Conference papers*:

  4. **Fabrizio Orlandi**, Pavan Kapanipathi, Amit Sheth, Alexandre Passant. *Characterising concepts of interest leveraging Linked Data and the Social Web.* In IEEE/WIC/ACM International Conference on Web Intelligence. Atlanta, GA, USA. 2013.

  5. **Fabrizio Orlandi**, John G. Breslin, Alexandre Passant. *Aggregated, interoperable and multi-domain user profiles for the Social Web.* In Proceedings of the 8th International Conference on Semantic Systems - I-SEMANTICS '12 (8). ACM Press 2012.

  6. **Fabrizio Orlandi**. *Multi-source provenance-aware user interest profiling on the Social Semantic Web.* In Proceedings of the Doctoral Consortium at User Mod-

eling, Adaptation and Personalization 20th International Conference, UMAP'12. LNCS Springer 2012.

7. **Fabrizio Orlandi**, Alexandre Passant. *Semantic search on heterogeneous wiki systems*. In Proceedings of WikiSym'10, International Symposium on Wikis. ACM Press 2010.

- *Workshop, demo and poster papers*:

8. Pavan Kapanipathi, **Fabrizio Orlandi**, Amit Sheth, Alexandre Passant. *Personalized filtering of the Twitter stream*. In 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation (SPIM 2011) at ISWC'11. CEUR-WS 2011.

9. **Fabrizio Orlandi**, Pierre-Antoine Champin, Alexandre Passant. *Semantic representation of provenance in Wikipedia*. In Second International Workshop on Role of Semantic Web in Provenance Management (SWPM 2010) at ISWC'10. CEUR-WS 2010.

Contents

Chapter 1

Introduction

1.1. Motivation and Problem Statement

The extraction, analysis and representation of information about users' social activities on the Web plays an important role for software systems providing personalisation and recommendations to their users. The demand for personalisation on social media websites, search engines, e-commerce websites, etc. is clearly growing and becoming an essential part of every relevant Web application. Popular examples are Amazon's product recommendations[1] and Google's targeted advertisement[2]. A challenge for Web application providers is to offer accurate personalisation without having to ask for users' explicit input or make users spend time on a long initial training period on the system (the typical "cold start" problem of recommender systems [Schein et al., 2002]). To overcome this challenge it is important to create accurate user models and, ideally, to aggregate relevant information about users from different sources on the Web [Carmagnola et al., 2011].

The process of mining user interests from Social Web platforms and representing them through accurate user profiles is crucial for any personalisation task. For example, for an e-commerce recommender system such as Amazon's, it is essential to collect and store information about the users' visited pages and acquired items. However, for such a system it would be extremely beneficial to gather additional information about its users' preferences expressed on other websites. In order to personalise users' experience on the Web (with their explicit consensus) it becomes necessary to deploy a

[1] http://www.amazon.com (accessed January 2014)
[2] Google AdSense: http://www.google.com/adsense (accessed January 2014)

methodology for retrieving possible concepts/entities[3] from different Social Web activities and selecting the most relevant ones which would match the real users' personal interests. This methodology should (i) adapt to different types of sources and online activities, (ii) model and represent personal structured collections, or profiles, of interests in an interoperable way and (iii) allow any application for different filtering and reasoning strategies over the concepts of interest depending on the personalisation task.

In this regard the *Web of Data*[4] is certainly a valid and extensive source of information for profiling and recommendation algorithms. The Web of Data offers a very large set of background knowledge in the form of structured data from different domains and communities publicly available on the Web. Popular examples range from the encyclopedic knowledge of DBpedia[5], to the DrugBank database[6] of drugs, to the musical artists and albums curated by the BBC[7]. The Web of Data provides easily accessible and machine readable data that can help with solving the "cold start" problem and enriching the level of detail of user profiles. Another solution that could help with this problem is the aggregation of user data from different social media sources. For example, if on one particular social platform not enough data about a user is available, in order to provide accurate personalisation, it would be possible to use the additional data contained in her aggregated multi-source user profile. Interoperability between social media websites is key in this context, where Semantic Web technologies could be adopted for the standard representation of the websites, their social activities and the user profiles.

Semantic technologies also play an important role in mining and selecting concepts and entities accurately representing users' interests. In this case, it is necessary to (i) distinguish between the different types of online user activities, (ii) recognise the important concepts of interests from user generated structured or unstructured data and (iii) select the most appropriate ones according to their semantics.

In this thesis we propose a methodology for profiling user interests that leverages semantic technologies for interlinking social websites and provenance management of

[3] In this thesis we adopt the following definitions of entities and concepts. An *entity* as a thing which is recognized as being capable of an independent existence and which can be uniquely identified. An entity is an abstraction from the complexities of a domain existing in the real world [Beynon-Davies, 2003]. Similarly, a *concept* is a fundamental category of existence, a mental representation which the brain uses to denote a class of things in the world. It is an entity that exists in the brain [Zalta, 2009].

[4] The Web of Data, as opposed to a traditional Web of documents, refers to a Web that can be processed directly or indirectly by machines (Section 2.1.2).

[5] `http://dbpedia.org` (accessed January 2014)

[6] `http://ckan.net/dataset/fu-berlin-drugbank` (accessed January 2014)

[7] `http://datahub.io/dataset/bbc-music` (accessed January 2014)

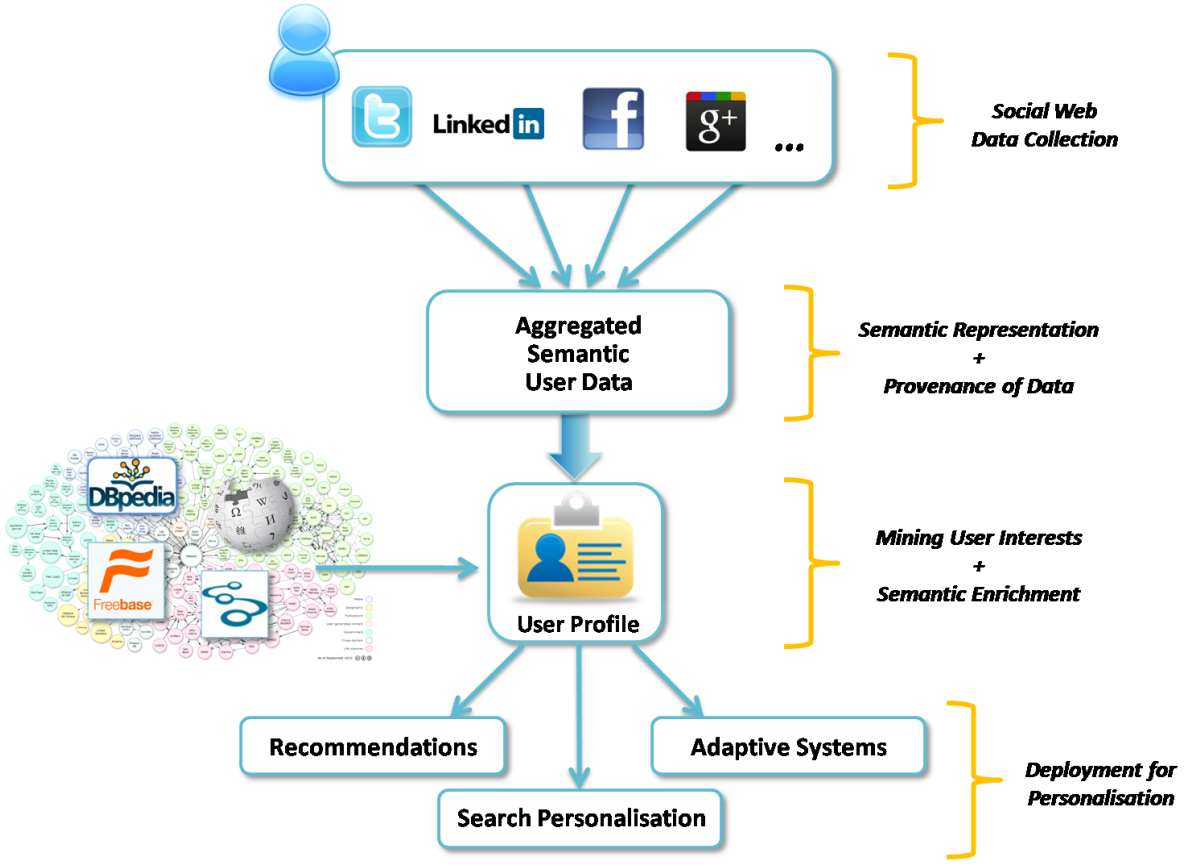

Figure 1.1.: Overview of the methodology for profiling user interests discussed in this thesis.

data on the Web to retrieve accurate information about data producers (either applications, software agents or users). We explore how provenance of data can be considered as one of the core building blocks for establishing data quality measures, for enhancing the knowledge acquisition/filtering process, and the user profiling phase. The goal is to build comprehensive profiles of user interests based on qualitative and quantitative measures about user activities across social sites. This would be possible by interlinking online communities using semantic technologies and popular lightweight ontologies, and by enriching the retrieved user data with information available on the Web of Data. Additionally, we investigate and evaluate a set of heuristics for mining users' interests from their social activities on heterogeneous social media websites and propose different approaches and measures for aggregating, enriching and ranking users' concepts of interest.

1.2. Research Questions

In the following chapters the thesis will address and answer several research questions, which are summarised in this section.

The current state of the Social Web, and especially of its websites offering personalisation, is composed of many different services targeting specific communities and offering particular products or applications. On each of these services users gain personalisation and recommendations by providing (implicitly or explicitly) information about themselves and in particular about their interests. However, all these social platforms hardly communicate with each other or allow users to reuse their own data in order to: improve personalisation, reduce time and effort in the creation of user profiles, or give users the ability to manage their own personal data. A step towards the solution of this problem is in developing a standard methodology for user profiling on the Social Web. Having this goal in mind, we can ask the following question, which is the core research question of the thesis:

> *How can we effectively collect, represent, aggregate, mine, enrich and deploy user profiles of interests on the Social Web for multi-source personalisation?*

The answer to this question would provide us a complete methodology for profiling user interests (Figure 1.1) that goes:

- from the collection and aggregation of user data from heterogeneous Social Web platforms,
- to the management and representation of this data,
- to the semantic enrichment of interoperable user profiles,
- to their adaptation and deployment for different personalisation tasks.

This thesis aims at identifying the main factors and challenges that influence user modelling and personalisation on the Social Web. In particular, we divide the main methodology and our investigation into three parts, leading to additional and more specific research questions (as depicted in Figure 1.2):

1. **Aggregation of Social Web data for profiling user interests**: *How can we aggregate and represent user data distributed across heterogeneous social media systems for profiling user interests?*

In particular, with this question we investigate how to leverage Semantic Web technologies for solving the challenge of aggregating different social media sites. As users on the Social Web interact with many different types of activities, it is necessary to capture this variety of Social Web actions. We analyse how to use these different kinds of activities and the user generated content in order to retrieve concepts of interest. Additionally, we study a possible formal representation of the retrieved user data for standardised and interoperable user profiles.

2. **Provenance of data for user profiling**: *What is the role of provenance on the Social Web and on the Web of Data and how to leverage its potential for user profiling?*

   We investigate how to select relevant features (and metadata) for user profiling from the collected user data according to its provenance. We identify the important interplay between the Social Web and the Web of Data through provenance information and describe the benefit of using provenance of data in the user profiling context. Recording and representing provenance at the stage of data collection and aggregation is beneficial to the enrichment and deployment stages of our profiling methodology. Therefore, the role of provenance of data on our methodology is transversal, it involves every stage of the profiling process.

3. **Semantic enrichment of user profiles and personalisation**: *How to combine data from the Social and Semantic Web for enriching user profiles of interests and deploying them to different personalisation tasks?*

   Following the collection and aggregation of user data and the semantic representation of the concepts of interest, it is necessary to select and filter user interests according to several measures and the particular use case for the profiles. Different use cases, or personalisation tasks (such as recommendations or user adaptive interfaces), require distinct types of concepts of interest and therefore distinct profiling strategies. Under this question, we investigate a number of strategies and measures for semantic enrichment of the concepts of interest. We employ the Web of Data and the Social Web for the enrichment and evaluate the performance of our measures in selected scenarios.

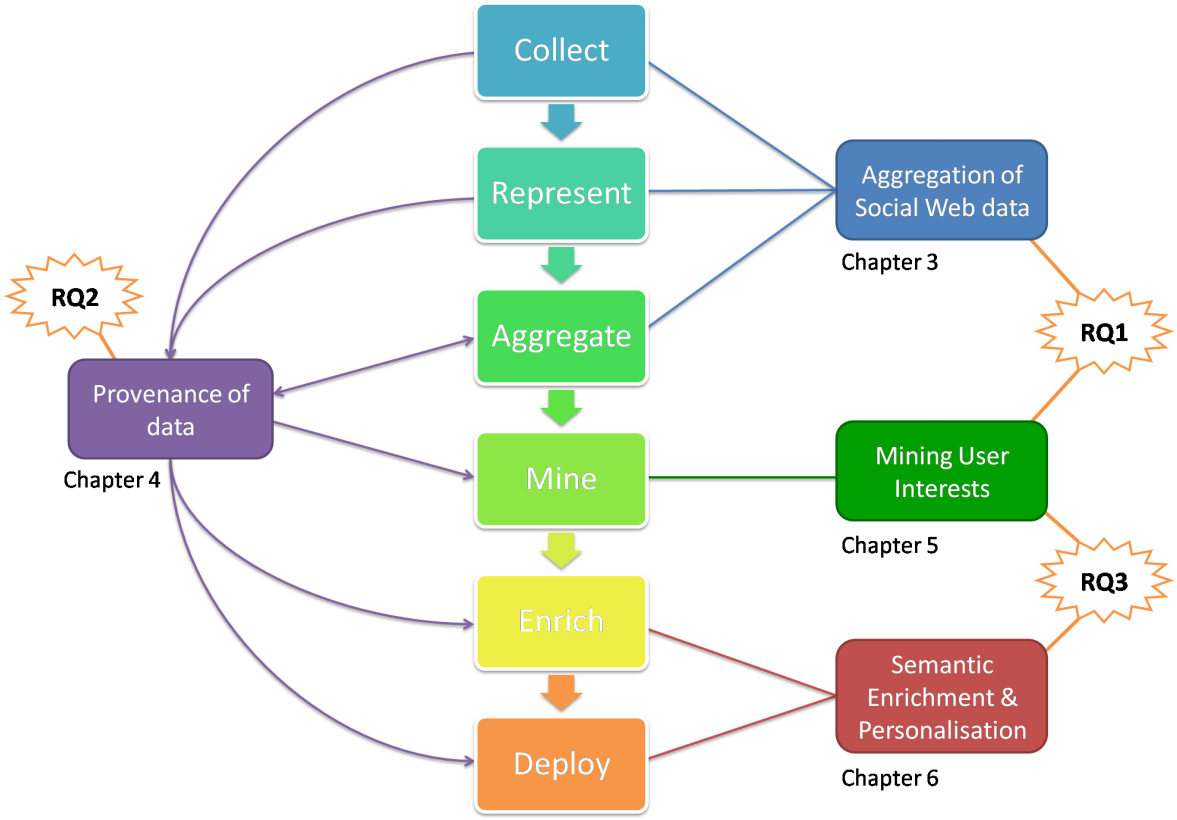

Figure 1.2.: The methodology for profiling user interests as formalised by the main research question (from "Collect" to "Deploy"), its connection with the three "sub-questions" (indicated with "RQx") and the chapters of the thesis covering them.

1.3. Thesis Overview

1.3.1. Research Map

In this section we provide a brief overview of our main contributions and the involved research areas, while more details on the implementation, the results and their evaluation are provided in the next chapters. As depicted in Figure 1.3, our contributions combine research in the areas of user modelling and personalisation, knowledge representation and provenance of data, Social and Semantic Web and information extraction.

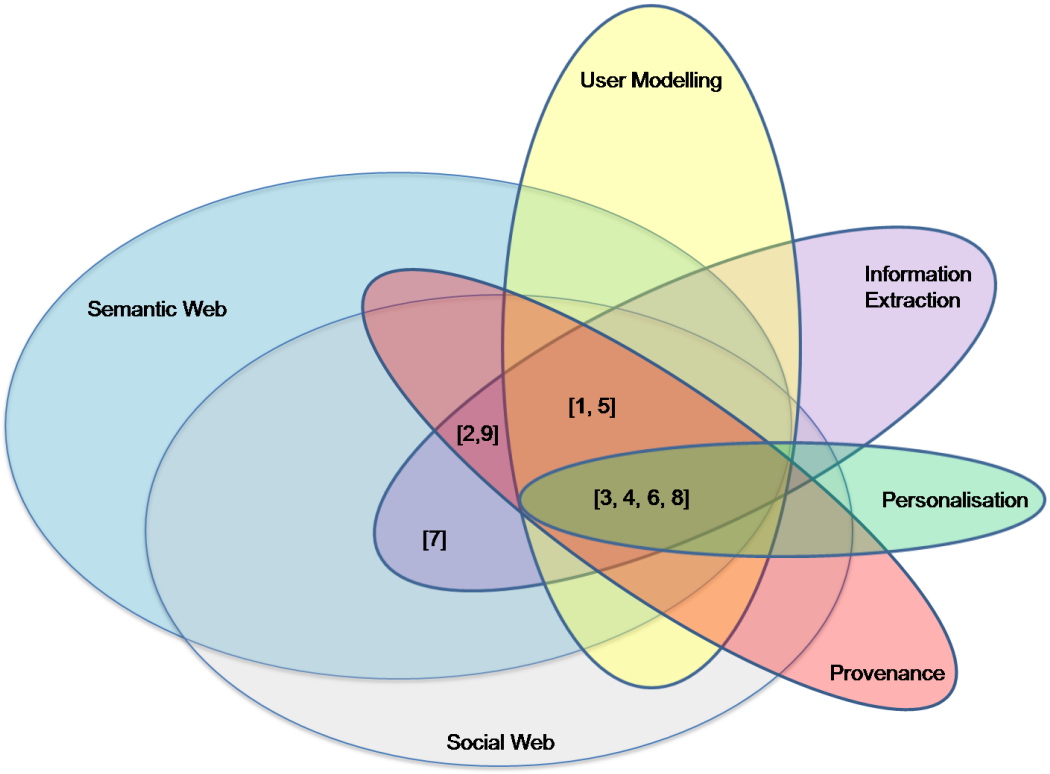

Figure 1.3.: Overview of the main research areas and methodologies covered by the thesis and our related publications (the numbers correspond to the ones indicated in the *Publications* section at the beginning of the thesis)

Contributions:

1. *A modelling solution for representing heterogeneous social media sites, their structure and their user activities facilitating the aggregation and analysis of distributed Social Web data.*

   This solution includes a characterisation of social media and users' Social Web activities and a study of the different machine readable vocabularies available for their representation. In this regard, we propose and extend several standard ontologies and, as an evaluation of the validity of the approach, a framework for the semantic representation and data management of wikis has been implemented. In particular, we describe an efficient application with a simple user interface enabling semantic searching and browsing capabilities on top of different interlinked wikis. More details are provided in Chapter 3 and our publication [Orlandi and Passant, 2010].

2. *An approach for modelling and managing provenance of data on both the Social Web and the Web of Data and use it as core element for mining user interests.*

In particular, we investigate how provenance of data can be extracted, represented and used on the Social Web. We demonstrate the importance of semantic representation and management of provenance and the close relationship between the Social Web and the Web of Data as connected by provenance information. More in detail, as a practical example, we provide a solution for representing and managing provenance of data from social media platforms such as wikis using Semantic Web technologies. In addition, we describe a specific lightweight ontology for representing provenance in wikis and a complete framework for the extraction of provenance data. An application for accessing the generated data in a meaningful way and exposing it to the Web of Data has been implemented and evaluated [Orlandi et al., 2010]. Moreover, we introduce an approach for modelling and managing provenance on the Web of Data using information extracted from the Social Web. This approach demonstrates the benefit of combining the Social Web and the Web of Data for understanding user interactions and preferences on social media sites. A modelling solution, an information extraction framework and a provenance-computation system have been implemented [Orlandi and Passant, 2011] (Chapter 4).

3. *A methodology and a set of heuristics for the creation and aggregation of multi-source user profiles of interests built on top of provenance information and aggregated/structured Social Web data.*

   We propose a methodology and a set of measures for: collecting user data from different social media sites, enriching the data with provenance information and Linked Data, and identifying and ranking relevant concepts of interest for user profiles. The methodology [Orlandi et al., 2012, Orlandi, 2012] has been implemented and evaluated with a system that aggregates user data from sites such as Twitter, Facebook, etc. and allows users to manage, and use, their user profiles of interests for personalisation. More details are provided in Chapter 5.

4. *A methodology for the semantic enrichment of user profiles of interests and its implementation for personalisation.*

   We propose a real-time, computationally inexpensive, domain independent model for characterising concepts of interest composed of: categorisation, popularity, temporal dynamics and specificity [Orlandi et al., 2013]. We describe and evaluate novel algorithms for computing these measures leveraging the semantics of Linked Data and evaluate the impact of our model on user profiles of interests (more de-

tails in Chapter 6) We detail our implementation of a Semantic Web approach to filter public microblog posts matching interests from personalised user profiles. Our approach includes automatic generation of multi-domain and personalised user profiles of interests, filtering Twitter stream based on the generated profiles and delivering them in real-time [Kapanipathi et al., 2011b, Orlandi et al., 2014].

Other contributions:

5. *A system that allows users to set fine-grained privacy preferences for the creation of privacy-aware faceted user profiles on the Social Web.*

   We implemented an architecture that provides users full control over their profile allowing them to define and show different facets of the profile based on fine-grained privacy preferences. The architecture allows for aggregation of profiles generated across different social websites (Facebook, LinkedIn, Twitter), scrutability of the profiles and management of access control rules [Sacco et al., 2012]. This contribution has been developed by merging our work on user profiling together with the work of Sacco et al. on privacy on the Social Semantic Web. This research, which is not a core contribution of the thesis, is only briefly summarised in Chapter 5 but raises awareness on privacy, an important aspect of user modelling and personalisation.

6. *A distributed real-time architecture for filtering and personalising large streams of messages on the Social Web.*

   As a result of the ongoing collaboration with the Kno.e.sis Centre[8] at Wright State University — in particular with Prof. Amit Sheth and Pavan Kapanipathi — we developed a novel solution for the personalisation of any public Social Web stream of messages in real-time. In particular we experimented our developed system with a real-time personalisation of the public Twitter stream. This contribution is partially described in Chapter 6, especially the aspects related to our research and expertise. The Kno.e.sis group mainly contributed with their expertise on the development of a scalable and distributed architecture capable of processing thousands of tweets per second [Kapanipathi et al., 2011b].

7. *International academic activities.*

[8] http://knoesis.wright.edu/ (accessed January 2014)

In addition to the papers published and presented at relevant conferences and journals in the field (see the *"Publications"* section at the beginning of this dissertation), we also engaged in other relevant academic activities. Among the most important ones we underline: the participation to the W3C Federated Social Web Incubator Group[9], the VIVO 2011 Conference[10] as invited expert and the SIOC project[11] as active contributor. Finally, we took part in the Organising Committee of the 2011 Web Science Summer School[12] and other relevant Programme Committees.

1.3.2. Document Structure

As illustrated in Figure 1.2, this dissertation is structured as follows. In this chapter, we introduced the research questions and provided an overview of the overall research plan, the motivations and the contributions. The following Chapter 2 presents a description of the state of the art related to the core research areas of the thesis. It introduces the most relevant definitions and the related work for the research fields of Social and Semantic Web, provenance of data, user modelling and Web personalisation.

As in Figure 1.2, Chapters from 3 to 6 are the core chapters of the thesis and include our main contributions. The order of the chapters follows the order of the steps for the user profiling methodology which we propose (from *Collect* to *Deploy*). Every chapter is dedicated to some phases of our profiling pipeline and, at the same time, focuses on specific research questions.

Chapter 3 provides the foundations of the methodology for profiling user interests. It describes a characterisation of social media and introduces our semantic modelling solution for representing Social Web sites and user activities. This semantic representation of social media is the necessary ingredient for creating a structured and interoperable meta-layer of Social Web data that can be used to aggregate user information and mine user interests. In this chapter we detail our model for social media that uses popular Semantic Web ontologies. Moreover, we describe a practical experiment that applies our model to a system integrating Social Web data from different heterogeneous sources. Hence, Chapter 3 describes the first steps for profiling user interests: from the collection of Social Web data to its semantic representation and aggregation. Aggregation and

[9] http://www.w3.org/2005/Incubator/federatedsocialweb/ (accessed January 2014)
[10] http://www.vivoweb.org/blog/2011/05/2011-vivo-hackathon-report (accessed January 2014)
[11] http://sioc-project.org/ (accessed January 2014)
[12] http://webscience.deri.ie/schools/2011/index.html (accessed January 2014)

mining of user interests on top of this structured Social Web data layer is described later in Chapter 5.

In Chapter 4 we describe how provenance of data plays a crucial role in social media and the Web of Data and especially for user profiling. In particular, we show how it can be recorded and represented on the Social Web and consequently used on the Linked Data cloud to track the origins of particular statements and data records. At the same time, provenance on/for the Web of Data can be used in many different use cases supporting Social Web users. Provenance of data is described as the fundamental connection between the Social Web and the Web of Data. It fuels with useful information every step of our profiling methodology (Figure 1.2).

In Chapter 5 we detail the core of our methodology for the automatic creation and aggregation of interoperable and multi-domain user profiles of interests. In particular, we describe how we mine and aggregate user interests extracted from social media data (following the steps defined in Chapter 3) along with its related provenance information (Chapter 4). Hence, we evaluate the effect of different provenance-based dimensions and heuristics on mining and ranking user interests in order to increase the accuracy of the user profiles. In this regard, a user study, conducted with Facebook and Twitter user accounts, is included in the chapter. We conclude Chapter 5 supporting the importance of privacy in our research and describing a management system for privacy preferences on user profile data.

While Chapter 5 (together with Chapter 3) concludes our analysis related to the first research question about aggregation of Social Web data for profiling user interests, it also starts our investigation about the third research question on semantic enrichment of user profiles and personalisation. This question is the main focus of Chapter 6. Here, we introduce a methodology for semantic enrichment and characterisation of concepts of interest. We employ the Web of Data and the Social Web for the enrichment and evaluate the impact of our measures on personalisation use cases. More in detail, we propose and evaluate a real-time, computationally inexpensive, domain independent model for concepts of interest. Then, we describe how to deploy enriched user profiles on practical personalisation use cases. In particular, we evaluate our complete profiling methodology on a personalisation system implemented for real-time filtering of Social Web streams of messages.

We conclude the thesis with Chapter 7, where we summarise the main results obtained and discuss lessons learned and possible future work. Our answers to the research questions are also outlined in this chapter as well as novel directions of research.

Chapter 2

Background

2.1. Towards the Social Semantic Web

2.1.1. Social Web

According to the World Wide Web Consortium (*W3C*), "the Social Web is a set of relationships that link together people over the Web"[1]. It consists of a combination of people and the Web, but it is not only about relationships between people, it is rather built around the connections between people and their objects of interest. This view of the Social Web, described in [Breslin et al., 2009] and originally introduced by sociologist Karen Knorr-Cetina [Knorr-Cetina, 1997], argues that the connections created by people on online social websites are established through "social objects" of common interest: *e.g.* the content they create together, co-annotate, or for which they use similar annotations. Therefore, what clearly distinguishes the Social Web from the traditional Web is the ability of users to interact with each other or with the content published on the Web.

2.1.1.1. Web 2.0

One of the fundamental changes of the Web in the early 2000s was a move from a consumer to a producer status of the users. Due to the introduction of new usage patterns and technologies, readers could react to the information they browsed in different ways.

[1] http://www.w3.org/2005/Incubator/socialweb/XGR-socialweb-20101206/ (accessed January 2014)

This technological change is commonly referred to as *Web 2.0*, a term initially coined by Darcy Di Nucci in 1999 [DiNucci, 1999] and later on made popular by Tim O'Reilly in 2005 [O'Reilly, 2005]. Although Web 2.0 is a very popular term, it is difficult to give its precise definition. It refers to a second generation of Web communities and services based on new structures and abstractions emerged on top of the ordinary Web. It is commonly perceived that Web 2.0 is the Web where people meet, collaborate and share anything that is interesting to them by using social software applications. Hence, the introduction of social aspects into Web 2.0 applications is a dominant factor.

In this regard, it could be appropriate to consider (as a "crowdsourced" definition) what current Web users say about Web 2.0. The encyclopedic definition from the related English Wikipedia[2] article is as follows:

"Web 2.0 describes Web sites that use technology beyond the static pages of earlier Web sites. [...] Although Web 2.0 suggests a new version of the World Wide Web, it does not refer to an update to any technical specification, but rather to cumulative changes in the way Web pages are made and used. A Web 2.0 site may allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated content in a virtual community, in contrast to websites where people are limited to the passive viewing of content. Examples of Web 2.0 include social networking sites, blogs, wikis, folksonomies, video sharing sites, hosted services, Web applications, and mashups. [...]" [3].

This definition is a perfect example of Web community cooperation through Web instruments provided by this evolution of the Web (In this case the instrument is Wikipedia, the popular wiki website).

As mentioned in the above definition, Web 2.0 led to a second generation of Internet-based services such as blogs, wikis, social media sites, communication tools and social networking services (SNS). In accordance with Tim O'Reilly [O'Reilly, 2005], the meaning of Web 2.0 can be presented by contrasting the traditional Web with the newer Web 2.0, as displayed in Figure 2.1.

Popular examples are Facebook[4] (currently the most popular SNS), Twitter[5] (a microblog), Wikipedia (an encyclopedic wiki), YouTube[6] (a social media, video-sharing,

[2] http://en.wikipedia.org (accessed January 2014)
[3] http://en.wikipedia.org/wiki/Web_2.0 (accessed January 2014)
[4] http://www.facebook.com (accessed January 2014)
[5] http://twitter.com (accessed January 2014)
[6] http://www.youtube.com (accessed January 2014)

Web 1.0		Web 2.0
DoubleClick	-->	Google AdSense
Ofoto	-->	Flickr
Akamai	-->	BitTorrent
mp3.com	-->	Napster
Britannica Online	-->	Wikipedia
personal websites	-->	blogging
evite	-->	upcoming.org and EVDB
domain name speculation	-->	search engine optimization
page views	-->	cost per click
screen scraping	-->	web services
publishing	-->	participation
content management systems	-->	wikis
directories (taxonomy)	-->	tagging ("folksonomy")
stickiness	-->	syndication

Figure 2.1.: From Web 1.0 to Web 2.0, as in [O'Reilly, 2005]

website), etc.[7] Web 2.0 applications derive from technologies such as Rich Internet Applications (RIA), Asynchronous JavaScript and XML (AJAX), Extensible HyperText Markup Language (XHTML), Cascading Style Sheets (CSS), Syndication and aggregation of data in RSS or Atom, clean and meaningful URLs. The introduction of these technologies allowed users of Web 2.0 to feel as if they used traditional desktop applications to share their content with the online communities.

It is important to note that there are two main principles constituting this evolution of the Web, as described in [Passant et al., 2009a].

- The first one is the *"Web as a platform"*, or the shift to the Web as the most important mean to deliver new services and applications. This implies the migration from traditional desktop applications (email clients, office suites, etc.) to Web-based applications.
- The second one is the *"architecture of participation"* principle, which represents how transparently each consumer becomes a data producer in Web applications based on the particular design of these services.

While the Web 2.0 could be seen as a technological wave of changes that contributed to the current Social Web, the principles behind online communities and online social networks define the sociological aspects of the current Web. As described in [Tapscott and Williams, 2006], the changes brought by the Web 2.0 were mainly sociological and economical, rather than technical. However, for a deeper understanding of the related development practices and Web 2.0 design patterns/principles we suggest the reader to consult [O'Reilly, 2005] and [Governor et al., 2009].

[7] To get a more comprehensive and updated list of Web 2.0 services the user could check `http://techcrunch.com` (accessed January 2014)

2.1.1.2. Online Communities

Online communities are groups of people that primarily interact via several different types of communication media (e.g. mobile phones, Internet, email, social network service on the Web, newsletter, etc.). The main reason why a user belongs to a social network is the desire to share and meet others with a similar domain of interests. Collaboration is a good way of reaching information and knowledge.

A social network service focuses on building online communities of people who share interests and/or activities, or who are interested in exploring the interests and activities of others. Most social network services are Web based and provide a variety of ways for users to interact, such as asynchronous messaging facilities and instant messaging services. Social networking has encouraged new ways to communicate and share information. Social networking websites are being used regularly by millions of people, and nowadays social networking is an enduring part of everyday life.

This definition corresponds to the definition of *communities of interest*, which are largely the most popular communities currently on the Web. However, there are also other types of communities, such as: communities of practice, communities of place, spontaneous and ephemeral communities, etc. In these communities users interact not only because of shared interests but also because of other reasons such as similar professional expertise, same location or shared participation to an event. In this thesis we focus mainly on communities of interest and we argue that our methodology is generic enough to be applicable also on other types of communities. After all, even in other types of communities users perform social activities which are related to particular personal interests. Hence, we can extend this thesis also to other types of communities.

Communication can be divided to three modes, classified on the basis of the techniques used: one-to-one (e.g. direct messages, etc.); one-to-many (e.g. Web pages, blogs, etc.); many-to-many (e.g. forum, wikis, etc.). Networks have diverse sizes. As an example, we depict a small social network in Figure 2.2 taken from [Diewald, 2012]. In a small, tight network, there are few people who form a kind of a private area. However, there can also be a lot of participants with loose connections (weak ties). From the collaboration point of view, the latter mode is more valuable as it is more probable to introduce new ideas [Granovetter, 1973]. Hence, it is better to have connections with other networks than with only one. However, unlimited access to information exchange can involve some risk: there is a possibility that a social network is flooded with un-

needed information. To avoid that, or at least to improve data/information quality, rating and annotation of shared resources were introduced.

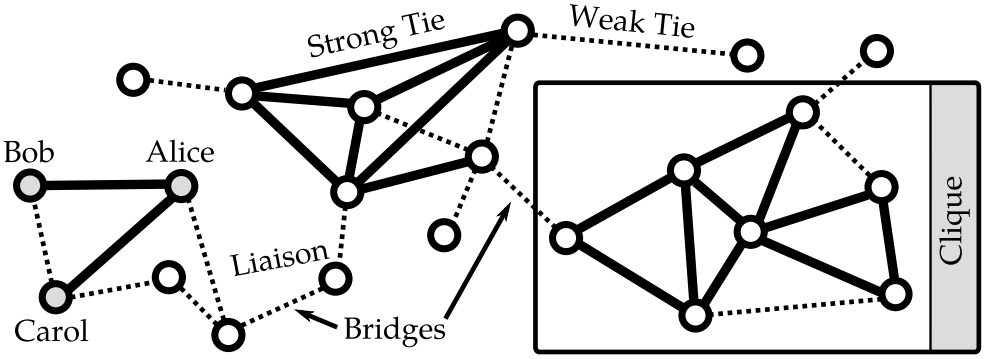

Figure 2.2.: Illustration of a small social network with three cliques connected via bridges. There are strong ties between the individuals Alice and Bob, and Alice and Carol. Based on [Granovetter, 1973], there is at least a weak tie between Bob and Carol. Granovetter defines ties as "a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie."

Social networks and online communities can be represented and studied using graph theory. The social connections between users, and their relations with online objects due to activities or interests, can be seen as edges (either typed or untyped, directed or undirected) of a graph where the vertices are the users and/or their objects of interaction (e.g. pictures, tags, comments, topics, etc.). These graphs can be analysed using traditional graph measures and modern social network analysis (SNA) techniques. Relevant related work which would provide an extensive and detailed introduction to social network analysis is in [Barabási, 2009] [Doreian and Everett,] [Wasserman and Faust, 1994] [Easley and Kleinberg, 2010] [Hanneman and Riddle, 2005]. While in [Erétéo et al., 2009] the authors propose a Semantic Web framework for describing and deploying SNA operators on any online social graph.

In Chapter 3 a precise characterisation of the main types of online communities and the related social media websites is provided.

2.1.2. Semantic Web

The Semantic Web is the ongoing evolution of the Web into a powerful and more reusable infrastructure for world wide information sharing and knowledge management. Initially

the Web, at least in its original and widespread version — first proposed by Berners-Lee in 1989 [Berners-Lee, 1989] then refined with some help from Robert Cailliau in 1990 [Berners-Lee and Cailliau, 1990] — was reduced to a very limited form. It mainly consisted of a publishing platform which allowed to connect with arbitrary information sources across all physical and technical boundaries; a publishing infrastructure of documents and links where very little consideration is given to the content or meaning of the documents or to the meaning of the links. In fact, Berners-Lee's 1989 proposal also envisioned a more expressive Web, where nodes and arcs on the Web (i.e. objects and relations) were not only limited to documents and untyped links respectively. Originally, it was proposed that both nodes and arcs on the Web could be arbitrarily typed to semantically represent anything. For instance, nodes could represent concepts, people, objects, etc. and arcs would represent particular relationships in between nodes, such as referral, dependencies, subsumption, etc. Thus, the Web was always conceived as a Web of typed resources semantically connected. However, its initial development and its most widespread form was very limited. It basically served as an excellent giant document repository and, as a communication platform, enabling the provision of various online services. Knowledge reuse was limited because no uniform standard was available to express the meaning or intended usage of pieces of online information.

Ten years after the WWW's conception, in 1999, Berners-Lee developed his original proposal of the Web further, naming it *Semantic Web* [Berners-Lee and Fischetti, 1999]. The Semantic Web [Berners-Lee et al., 2001] is a Web of information that is more understandable and more usable by machines than the current Web. To use its author's words, it is where computers "become capable of analysing all the data on the Web - the content, links, and transactions between people and computers". It can be regarded as an extension of the existing Web, whose information is mostly human-readable. Although the current Web also has some machine-usable structure such as head and body of documents, levels of heading elements, classes of `<div>` elements[8], this structure has coarse granularity and little agreed-upon meaning. The Semantic Web allows for finer granularity of machine-readable information and offers mechanisms to reuse meaning. It can also be considered similar to a large online database, containing structured information that can be queried. But in contrast to traditional databases, the information can be heterogeneous: it does not conform to one single schema; the information can be contradicting: not all facts need to be consistent; the information can be incomplete:

[8] `<div>` is an HTML tag which expresses a block-level logical division.

not all facts need to be known; and resources have global identifiers allowing interlinked statements to form a global "Web of Data"[9].

In order to enable an interoperable and usable Web of Data, the Semantic Web relies on two main broad requirements:

- a common model to define Web resources and represent assertions about these resources. This is possible thanks to URIs - Uniform Resource Identifiers [Berners-Lee et al., 2005] - and RDF - Resource Description Framework [Klyne and Carroll, 2004a]- (Section 2.1.2.1);
- formal vocabularies to represent the semantics of Web resources and their assertions in an interoperable way. This is possible thanks to ontologies [Gruber, 1993] (Section 2.1.2.2), which can be defined for instance using RDFS - RDF Schema [Brickley and Guha, 2004] - and OWL - Web Ontology Language [Patel-Schneider et al., 2004].

The aforementioned set of technologies, which allows for a machine readable Web, has been brought forward mainly by the Semantic Web initiative, led by W3C since 2001. W3C started a new activity in December 2013, called Data Activity[10], that now subsumes the original Semantic Web Activity. The new Data Activity merges and builds upon the eGovernment and Semantic Web Activities. It aims at facilitating data publication on the Web continuing the previous effort led by the Semantic Web initiative. Working groups and standardisation efforts, conducted in the past decade by the W3C, developed into a complete novel high-level architecture, as extension to the hypertext Web. In Figure 2.3 we can see the *Semantic Web Stack* (also referred to as the Semantic Web Layer Cake) which illustrates the architecture of the Semantic Web.

The stack is a bottom-up sequence of standards and technologies based upon established hypertext technologies such as URIs as identifiers and UNICODE as character set. RDF and RDFS (often referred to as RDF/S or RDF(S) for brevity) are the core of the Semantic Web, representing the framework for data interchange and the basic vocabulary required to create taxonomies respectively (as we will detail later in this chapter). They are based on a syntax layer often represented by XML, but other serialisations can also be used. Ontologies and more expressive vocabularies are modelled with RDF(S)/OWL, while querying and storing the data can be done with SPARQL and

[9] "The Semantic Web is a web of data", as defined by the W3C: http://www.w3.org/2001/sw/ (accessed January 2014)

[10] http://www.w3.org/2013/data/ (accessed January 2014)

dedicated RDF stores (Section 2.1.2.3). Artificial Intelligence [Russell and Norvig, 2003] concepts such as formal logic, proof and trust are added as layers to the top of the stack, providing Semantic Web applications with additional complex inferencing capabilities.

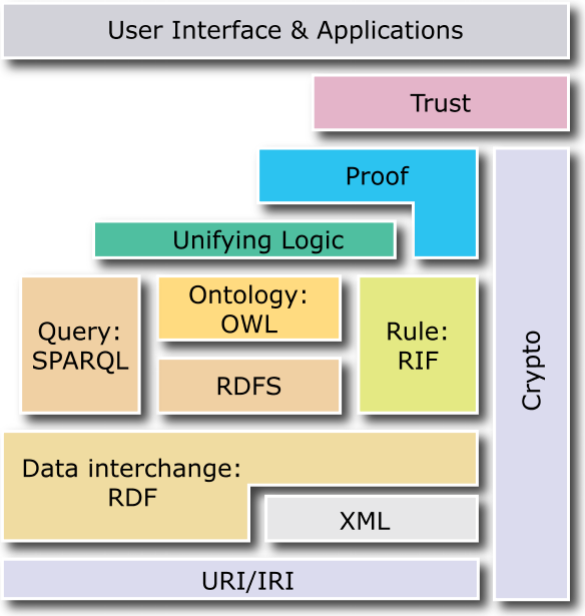

Figure 2.3.: The Semantic Web Stack - W3C[11]

In the following sections we will go deeper into the aforementioned core Semantic Web technologies and will have an overview of the main standards relevant to the work done in this thesis.

2.1.2.1. The Resource Description Framework, RDF

The fundamental data-model of the Semantic Web is the Resource Description Framework[12] (RDF) [Klyne and Carroll, 2004b]. RDF is a language for asserting statements about arbitrary identifiable resources. The use of global identifiers (URIs) [Ayers and Völkel, 2008] allows statements from different sources to interlink, ultimately forming a hypergraph of statements. For instance, URLs used on the Web (usually starting with `http:`) are a particular kind of URIs [Hansen et al., 2006] and, in fact, it is common and recommended to use `http:` URIs as identifiers. They can represent not

[11] from http://www.w3.org/2001/sw/ (accessed January 2014)
[12] http://www.w3.org/RDF/ (accessed January 2014)

only Web resources but also real world concepts and entities. They are the base of the Semantic Web and they should adhere to specific syntax and guidelines[13].

Although originally created to describe resources on the Web, such as pages and other content, RDF is domain-independent and can be used to model any information resource, world object, or abstract concept. RDF is a formal language in the sense that syntax, grammar, and model-theoretic semantics are defined [rdf, 2004]. It is a W3C standard and it was designed to be read and processed by machines, hence not to be displayed to people. It is a very primitive modelling language, however, more complex languages such as OWL are built on top of it.

In RDF there are only two types of primitives: resources and literals. Resources represent, ideally, everything that can be identified and described. They can be either identified by a URI or a *blank node*. Blank nodes are resources with a proper identity which, however, have an unknown or irrelevant identifier. On the other hand, literals are simply strings, or character sequences, without an identifier. They are used to specify a value or a description and eventually they can be associated with a language or a datatype identifier.

RDF is based on statement concepts. In a statement there is a *subject*, a *predicate* and an *object*; altogether they are called a *triple* (a statement). A collection of RDF statements produces a directed graph in which arrows point from subjects to objects whereas labels on arrows represent predicates. Subjects can be either URIs or blank nodes, predicates must be URIs and objects can be either URIs, blank nodes, or literals.

If we consider for example the following sentence: "*Fabrizio knows Alex*", it can be represented by an RDF statement that has the following structure: there is a subject (resource) *Fabrizio*, a predicate (property) *knows*, an object (value) *Alex*. Supposing all three parts are attributed with an URI, all with the namespace `http://example.com/` abbreviated with `example:` (QName), and that the names "Fabrizio" and "Alex" represent specific persons identifiable with a URI (i.e. the URI for Alex would be `http://example.com/Alex` or abbreviated, using the namespace, in `example:Alex`). Then, the above statement can be modelled in RDF and illustrated by a graph, as showed in Figure 2.4.

[13] In this regard we suggest the reader to consult the W3C Note "Cool URIs for the Semantic Web", December 2008, `http://www.w3.org/TR/cooluris/` and the original article by Berners-Lee "Cool URIs don't change", 1998, available at `http://www.w3.org/Provider/Style/URI` (accessed January 2014).

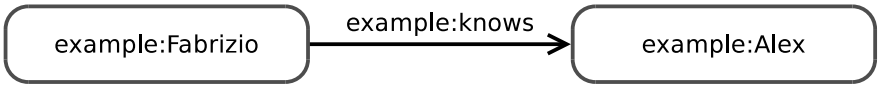

Figure 2.4.: RDF statement representing "Fabrizio knows Alex".

Besides the graph, a RDF serialisation, such as RDF/XML, can be used to show triples and relationships between them (see Listing 2.1). RDF/XML is an XML-based notation standardised by the W3C[14] and it is one of the most widely used syntaxes. To note that RDF/XML is only one of the multiple possible serializations for RDF data. Other serialisations, such as Turtle[15] or JSON-LD[16], are available and offer different advantages or disadvantages, depending on the use case. As we can see in Listing 2.1, the representation of a simple statement such as the one in Figure 2.4 is not easily readable. However, RDF/XML can be easily parsed by widespread XML tools.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:example="http://example.com/">
   <rdf:Description rdf:about="http://example.com/Fabrizio">
      <example:knows rdf:resource="http://example.com/Alex" />
   </rdf:Description>
</rdf:RDF>
```

Listing 2.1: RDF/XML representation of the statement in Figure 2.4

To make things a bit more interesting, we can add some other triples to the previous example to specify some additional information related to the mentioned entities. For example, we can state that the two resources identified by the URIs `http://example.com/Fabrizio` and `http://example.com/Alex` represent persons, and their names are "Fabrizio Orlandi" and "Alexandre Passant" respectively. This addition to the original statement is depicted as a graph in Figure 2.5 and modelled in Turtle language as in Listing 2.2.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix example: <http://example.com/> .

example:Fabrizio foaf:knows example:Alex .
example:Fabrizio rdf:type foaf:Person .
example:Alex rdf:type foaf:Person .
```

[14] RDF/XML Syntax Specification - W3C Recommendation - 2004 - http://www.w3.org/TR/rdf-syntax-grammar/ (accessed January 2014)

[15] Turtle: Terse RDF Triple Language - W3C Candidate Recommendation - 2013 - http://www.w3.org/TR/turtle/ (accessed January 2014)

[16] JSON-LD 1.0: A JSON-based Serialization for Linked Data - W3C Candidate Recommendation - 2013 - http://www.w3.org/TR/json-ld/ (accessed January 2014)

```
example:Fabrizio foaf:name "Fabrizio Orlandi"^^xsd:string .
example:Alex foaf:name "Alexandre Passant"^^xsd:string .
```

Listing 2.2: Extension and Turtle representation of the example in Figure 2.4. Four statements are added to the original example and real popular vocabularies are used (i.e. "foaf" and "rdf") instead of the generic "example" namespace.

As we can see in Listing 2.2, even in Turtle notation it is possible to use namespaces to abbreviate URIs: a feature available also in many other languages. Namespaces are identified by the `@prefix` string at the beginning of a statement. In this example we use an additional namespace, *FOAF*, whose resources reside at `http://xmlns.com/foaf/0.1/`. FOAF (Friend-of-a-Friend), which will be described more in detail later, is a popular project that published on the Web a specific RDF vocabulary (or ontology) describing persons, their activities and their relations to other people and objects[17]. By reusing the terms defined by FOAF and residing at the aforementioned namespace, we are able to use entities, terms, and concepts defined by a large community of experts. We also benefit from higher interoperability for our RDF data (see Section 2.1.2.4).

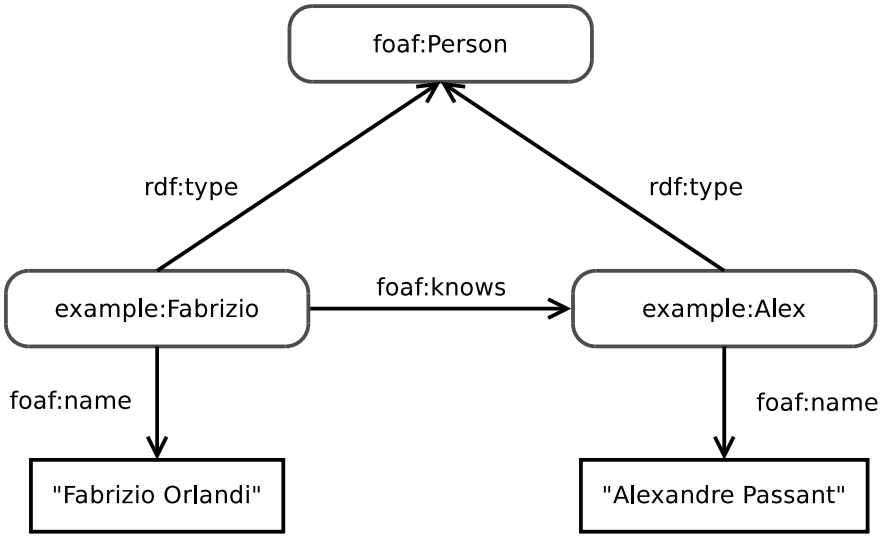

Figure 2.5.: Graph of the RDF document example depicted in Listing 2.2

RDF itself provides also a few fundamental terms for describing resources. For example it is possible to assign a type to a resource using the `rdf:type` property; as in our example assigning the type `foaf:Person` to the subject URIs `example:Fabrizio` and `example:Alex`. Further, the property `foaf:name` connects the URIs of the two persons in the example with their names as literals.

[17] `http://www.foaf-project.org/` (accessed January 2014)

The semantics described with RDF provide a formal meaning to a set of statements through an interpretation function into the domain of discourse. But this interpretation function is relatively straightforward and explicit: the semantics of RDF [Hayes, 2004] prescribe relatively few inferences to be made from given statements – there is only little implicit information in statements. RDF can thus be seen as a language for statements without specifying the meaning of these statements.

Such "machine-readable" meaning can be achieved by defining a vocabulary (a set of terms) for RDF and by specifying what should be done when such a term is encountered. Currently, two such vocabularies have been agreed upon and standardized: RDF Schema (RDFS) and Web Ontology Language (OWL).

RDF Schema (RDFS)[18], allows the expression of a schema-level information such as class membership, sub-class hierarchies, class attributes (properties), and sub-property hierarchies [RDF, 2004]. RDFS allows simple schema information, but its expressiveness is limited.

The Web Ontology Language (OWL)[19] extends RDFS (although the two are formally not completely layered) and provides terms with additional expressiveness and meaning [owl, 2004]. OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

In the next section "Vocabularies and Ontologies" a more detailed description of RDFS and OWL for defining ontologies is provided. Later, in Section 2.1.2.3, an overview of SPARQL, the standard language for querying and storing RDF data, is described.

2.1.2.2. Vocabularies and Ontologies: RDFS and OWL

The term ontology has its origin in philosophy, and has been applied in many different ways. In computer science and information science, an ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts. A widely cited paper [Gruber, 1993], associated with the effort to define this term, is credited with a formal definition of ontology as a technical term in computer science.

[18] RDF Schema 1.1. W3C Recommendation 25 February 2014: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/ (accessed April 2014)
[19] http://www.w3.org/2004/OWL/ (accessed January 2014)

An ontology is defined as an "explicit specification of a conceptualization," which is "objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold among them". In other words, it is an "abstract, simplified view of the world that we wish to represent for some purpose" [Gruber, 1993]. While the terms specification and conceptualization have caused much debate within the research community, the essential points of this definition of ontology are:

- An ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modelling a domain.
- The specification takes the form of the definitions of representational vocabulary (classes, relations, and so forth), which provides *meaning* for the vocabulary and *formal* constraints on its coherent use.

It is important to note that, since the purpose of the ontologies is to be used for knowledge exchange and integration and collaborative online work, ontologies on the Semantic Web should be *shared* and *agreed* upon by groups of users. Hence, we can state that an ontology on the Semantic Web is a shared and formal conceptualisation of a domain of discourse. This social aspect of ontologies is also demonstrated by the recent increase in development and adoption of lightweight ontologies shared among larger and larger communities. This is, for example, the case of *schema.org*[20], the Facebook Open Graph Protocol[21] and SIOC[22] as described in [Bojars, 2009].

In the technology stack of the Semantic Web standards (see "Layer Cake" in Figure 2.3), ontologies are called out as an explicit layer. There are now standard languages and a variety of commercial and open source tools for creating and working with ontologies. The standard formal ontology languages are RDF Schema (RDFS) [Brickley and Guha, 2004] and the Web Ontology Language (OWL) [Dean and Schreiber, 2004] (and its second edition "OWL 2", now a W3C Recommendation [W3C, 2012]). One of the main differences between RDFS and OWL is in their expressive power, which is higher in OWL. In RDFS it is possible to define lightweight *vocabularies*, while more complex *ontologies* are defined using OWL or other similar languages.

Typical elements of ontologies are in general: *concepts*, *properties* and *axioms*. While concepts, or classes, are either abstract or concrete objects of a particular domain, properties are the relationships between those classes and/or their instances. Like-

[20] http://schema.org/ (accessed January 2014)
[21] http://ogp.me/ (accessed January 2014)
[22] http://sioc-project.org/ (accessed January 2014)

wise, axioms define the logical assertions about the two aforementioned elements. Furthermore, it is important to distinguish between the ontology itself and its individuals or instances. The latter are not part of the conceptual model (the ontology) which is defining them but, together with the ontology, form the so called *knowledge base* [Guarino and Giaretta, 1995]. Similarly, in Description Logics [Baader et al., 2010], there is a difference between ABox and TBox, the first being the set of assertions and the second representing the model and the axioms.

In RDF terms, an ontology is the formal definition of classes, properties and instances used in a graph. RDFS extends RDF semantics with the following capabilities: definition of classes (using `rdfs:Class`), organisation of classes within hierarchies (`rdfs:subClassOf`), definition of domain ("subject") and range ("object") of properties (using `rdfs:domain` and `rdfs:range` together with `rdf:Property`), and organisation of properties within hierarchies (`rdfs:subPropertyOf`). Other properties designed for the human-readable annotation of resources, are introduced with RDFS, *i.e.* `rdfs:comment`, `rdfs:label` and `rdfs:seeAlso`. RDFS terms are used for the definition of all the RDF-based ontologies and vocabularies, even for OWL and RDFS itself. In its simplicity, this demonstrates the importance and flexibility of RDFS.

RDFS formal semantics [Hayes, 2004] are defined as a set of entailment rules and axioms. Thanks to these rules it is possible to use inference and entail additional statements over an existing graph of statements. A common example of RDFS inference rules is the *subsumption* of the `rdfs:subClassOf` and `rdfs:subPropertyOf` properties, as displayed in Table 2.1. All the RDF(S) rules are described in the W3C "*RDF Semantics*" Recommendation [Hayes, 2004]. The rule *rdfs9* in Table 2.1, for example, indicates that every instance `vvv` of a class `uuu` is also an instance of the super-class `xxx` of `uuu`.

Rule	**If**	**Then**
rdfs7	`aaa rdfs:subPropertyOf bbb .` `uuu aaa yyy .`	`uuu bbb yyy .`
rdfs9	`uuu rdfs:subClassOf xxx .` `vvv rdf:type uuu .`	`vvv rdf:type xxx .`

Table 2.1.: Example of RDFS inference rules [Hayes, 2004]: subsumption of properties and classes.

Although RDF and RDF Schema are helpful in expressing simple statements, they lack when used in more complex cases. That is why Web Ontology Language (OWL) was developed. It originates from the previous work on DAML+OIL [Horrocks, 2002] and

it became a W3C Recommendation in 2004 [Dean and Schreiber, 2004] after the W3C effort started in 2001 with a OWL Working Group. A second edition, named "OWL 2", became later[23] a Recommendation [W3C, 2012].

OWL consists of three sub-languages: OWL Lite, OWL DL and OWL Full. Each sub-language encapsulates the former ones. It is mainly their level of restrictions which distinguishes them. OWL *Full* supports the complete vocabulary without restrictions. OWL *DL* ("Description Logic") defines some restrictions (*e.g.* it imposes disjointness of classes, instances and properties) and OWL *Lite* restricts the available OWL vocabulary and imposes further restrictions on its use. With OWL 2, in addition to OWL 2 DL and OWL 2 Full, three additional *profiles* are specified: OWL 2 EL, OWL 2 RL, and OWL 2 QL. These additional profiles are designed to be approachable subsets of OWL 2 sufficient for a variety of applications[24].

In general, in addition to the capabilities offered by RDFS, OWL (and OWL 2) introduces:

(i) a new top level class (`owl:Class` subclass of `rdfs:Class`),

(ii) an extended vocabulary to define classes including enumeration of resources or union, intersection, complement of classes,

(iii) new specific classes for properties (`owl:DatatypeProperty` and `owl:ObjectProperty` subclasses of `rdf:Property`),

(iv) a new vocabulary to define inverse, functional, transitive or symmetric properties,

(v) an additional vocabulary for the annotation of ontologies and instances.

Hence, OWL allows not only the definition of ontologies with classes, properties and their instances. It also allows us to define cardinality constraints on properties, specifying transitivity, uniqueness, etc.

Similarly to RDFS, the axioms described in OWL can be used for reasoning. A reasoner, is a tool able to infer logical consequences from a set of axioms or assertions. Many different tools have been implemented for reasoning over OWL axioms and the differ on their conformance to standards, licensing , expressivity, reasoning algorithm, etc.[25]

[23] first in 2009 and then with a second release in 2012

[24] http://www.w3.org/2007/OWL/wiki/Primer#ref-owl-2-profiles (accessed January 2014)

[25] W3C list of OWL Implementations and Reasoners: http://www.w3.org/2001/sw/wiki/index.php?title=OWL/Implementations&oldid=3975 (accessed January 2014)

Despite being instrumental in advancing the Semantic Web, the first OWL standard raised a number of concerns [Grau et al., 2008]. In particular, inconsistencies between the interpretation of the different syntaxes, limitations of expressivity, and other problems with data types and RDF semantics have been identified by the community. For this reason the new OWL 2 version has been proposed and standardised [W3C, 2012]. The new Recommendation includes a metamodel based on the Meta Object Facility[26] (MOF), which addresses the inconsistency problem of the different syntaxes of OWL. This also eases the development of OWL APIs by increasing the interoperability of OWL 2. Solutions for the improvement of OWL's expressivity are also included with the addition of new useful features, *i.e.* qualified number restrictions, propagation of properties, richer data typing and keys for named individuals ("easy keys").

2.1.2.3. Querying on the Semantic Web: SPARQL

SPARQL is "a set of specifications that provide languages and protocols to query and manipulate RDF graph content on the Web or in an RDF store" [Harris and Seaborne, 2013]. Among the most important specifications it includes: a query language, an update language for RDF graphs, protocols for the execution of distributed queries, query results formats and entailment regimes. In particular, the *SPARQL Query Language* for RDF can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. Its name is a recursive acronym that stands for SPARQL Protocol and RDF Query Language. It is considered as the SQL of the Semantic Web and according to Tim Berners-Lee: "Trying to use the Semantic Web without SPARQL is like trying to use a relational database without SQL"[27]. It has been standardized by the SPARQL W3C Working Group[28] (was RDF Data Access Working Group) and on 15 January 2008, SPARQL 1.0 became an official W3C Recommendation[29], while the newer version SPARQL 1.1 in March 2013[30] [Prud'hommeaux and Seaborne, 2008].

SPARQL uses a graph pattern matching approach applied to graph data described with RDF. It provides capabilities for querying multiple required and optional graph patterns along with their conjunctions and disjunctions. Complex queries may include

[26] http://www.omg.org/mof/ (accessed January 2014)
[27] http://www.w3.org/2007/12/sparql-pressrelease (accessed January 2014)
[28] http://www.w3.org/2011/05/sparql-charter (accessed January 2014)
[29] http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/ (accessed January 2014)
[30] http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/ (accessed January 2014)

union, optional query parts, and filters. Moreover, new features such as value aggregation, path expressions and nested queries have been added in SPARQL 1.1. It also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be result sets or RDF graphs and different formats are supported (XML, JSON, CSV, TSV). In addition to SELECT queries, SPARQL supports ASK queries (i.e. boolean "yes/no" queries) and CONSTRUCT queries (allowing the creation of new RDF graphs from query results). Finally, DESCRIBE queries allow to obtain a graph describing a queried resource.

As an example, the following simple SELECT query returns *"people who were born in Dublin before 1900, ordered by name"*.

```
PREFIX exmpl: <http://example.com/exampleOntology#>

SELECT ?name ?birth WHERE {
   ?person exmpl:birthPlace <http://example.com/resource/Dublin> .
   ?person exmpl:name ?name .
   ?person exmpl:birthDate ?birth .
     FILTER (?birth < "1900-01-01"^^xsd:date) .
}
ORDER BY ?name
```

Listing 2.3: Example of a SPARQL query

In Listing 2.3 variables are indicated by a "?", and bindings for `?name` and `?birth` will be returned ordered by name (`ORDER BY` clause). The SPARQL `FILTER` clause restricts solutions to those for which the filter expression evaluates to TRUE, in this case if the `?birth` variable has a date value minor than "1900-01-01". The SPARQL query processor will search for sets of triples that match these four triple patterns, binding the variables in the query to the corresponding parts of each triple. To make queries concise, SPARQL allows the definition of prefixes and base URIs in a fashion similar to the Turtle RDF syntax (Section 2.1.2.1). In the query example above, the prefix `exmpl` stands for `http://example.com/exampleOntology#`.

An interesting feature added with SPARQL 1.1 is the ability to specify and execute updates to RDF graphs in a Graph Store[31]. This feature, called SPARQL Update[32], allows for insertion and deletion of triples and also load, copy and deletion of graphs through the use of simple queries.

[31] Here we adopt the W3C definition of Graph Store, which is a mutable repository of RDF graphs managed by one or more services

[32] `http://www.w3.org/TR/sparql11-update/` (accessed January 2014)

2.1.2.4. Linked Data

One of the main challenges faced during the early development of the Semantic Web was that it was generally designed by experts and not by regular Web users. Initially, the main focus of the Semantic Web community has been on the theoretical foundations of ontologies, data modelling, logic and reasoning. However, the need for a consistent and useful amount of Semantic Web data became more dominant than its modelling and processing. This is when, in 2007, the term Linked Data emerged. Linked Data refers to the methods used to expose, share and interlink structured data on the Web. Large datasets on the Web would then become more useful by publishing and/or interlinking them using open standard formats. To make this possible, Berners-Lee defined four principles for Linked Data [Berners-Lee, 2006b]:

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

The goal is then to build a large-scale graph of interconnected, open and structured data on the Web [Bizer and al., 2009]. The Linking Open Data (LOD) initiative, started in June 2007 by the Semantic Web Education and Outreach (SWEO) Interest Group[33], supported this vision with the publication of an impressive number of interconnected datasets in RDF openly on the Web. In January 2014 this number reached almost 62 billion RDF triples, from more than 2100 datasets, according to the *LODStats*[34] Web application constantly monitoring the LOD cloud [Demter et al., 2012] (see Figure 2.6).

In Figure 2.6 the most recent image representing the datasets part of the Linking Open Data cloud is displayed. It shows a diverse set of data sources which range from encyclopedic knowledge — such as DBpedia[36], the RDF export of Wikipedia [Auer et al., 2007] [Bizer et al., 2009] — to biomedical information [Jentzsch et al., 2009], to BBC music, news and TV programs [Kobilarov et al., 2009]. Most of the datasets are

[33] http://www.w3.org/blog/SWEO/ (accessed January 2014)
[34] http://stats.lod2.eu/ (accessed January 2014)
[35] by Richard Cyganiak and Anja Jentzsch (CC-BY-SA). http://lod-cloud.net/ (accessed January 2014)
[36] http://dbpedia.org (accessed January 2014)

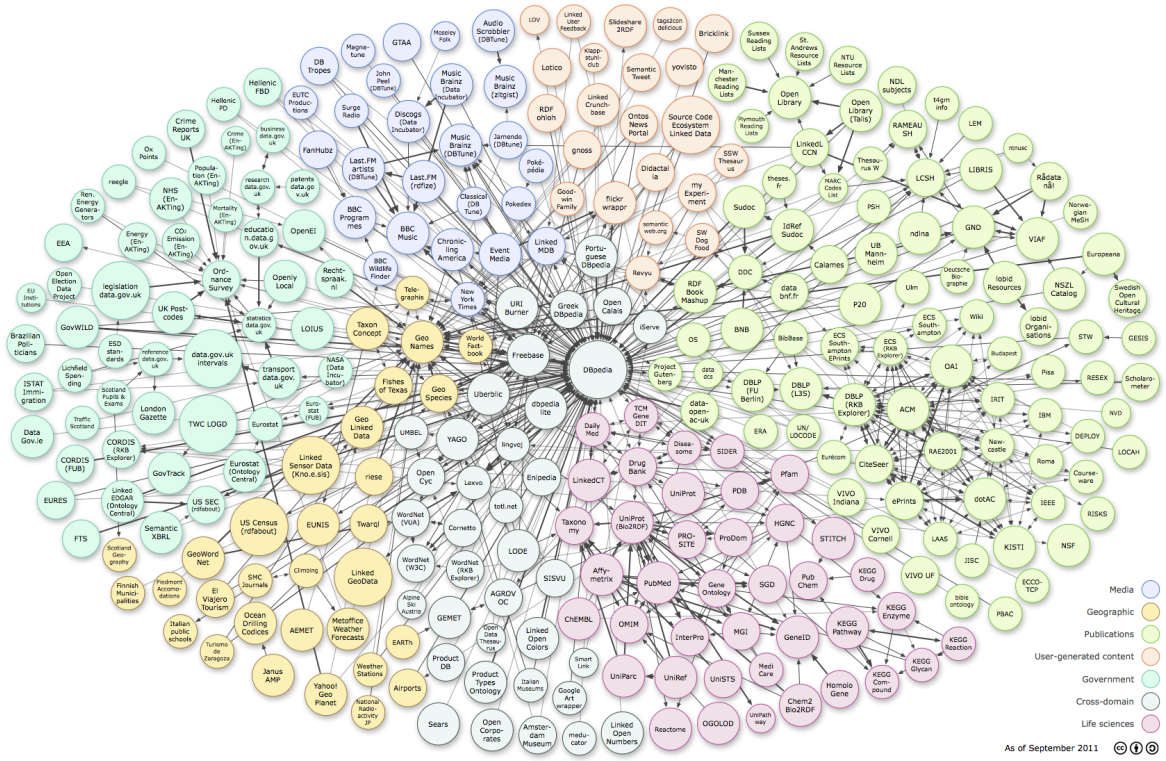

Figure 2.6.: Linking Open Data cloud diagram (September 2011)[35]

stored in particular databases designed for managing triples, called RDF stores or triple-stores. They usually expose a SPARQL endpoint (or a user interface on top of it) for querying and exploring the dataset.

The success of the Linked Data initiative is also dependent on the use of popular and shared lightweight vocabularies and ontologies. Important and widely used examples of ontologies publicly available on the Web are listed below. These are the most popular RDFS/OWL ontologies for specific domains of interests and are frequently mentioned and used in this thesis:

- people and social networks: Friend Of A Friend (FOAF)[37]
- online communities and discussions: Semantically-Interlinked Online Communities (SIOC)[38]
- documents: Dublin Core (DC)[39]

[37] http://www.foaf-project.org/ (accessed January 2014)
[38] http://www.sioc-project.org (accessed January 2014)
[39] http://dublincore.org/ (accessed January 2014)

- thesauri, taxonomies and subject-heading systems: Simple Knowledge Organization System (SKOS)[40]

In addition to these popular ontologies, we have to mention other widespread schemas and projects that have been adopted in popular and real consumer products. This adoption shows the success of Semantic Web technologies also in relevant business oriented projects. Facebook introduced the Open Graph Protocol[41], which is used on Facebook to allow any Web page to become a rich object in the Facebook social graph. Google adopted the Knowledge Graph[42], a structured knowledge base derived from the integration of many sources, including the CIA World Factbook[43], Freebase[44], and Wikipedia. It is currently used by Google to enhance its search engine's search results with additional structured information. Moreover, this is related to the introduction of Schema.org[45] a collection of schemas developed and adopted by the current most popular search engines of Bing, Google, Yahoo! and Yandex. This vocabulary can be used by webmasters to markup their pages in ways recognized by major search providers. A mapping from the terms defined in Schema.org to RDF (expressed in RDF Schema) has been created by the Linked Data community[46].

2.1.3. Social Semantic Web

Not all the concepts belonging to the original view of the Web, as described by Berners-Lee in 1990 [Berners-Lee and Cailliau, 1990], were brought to fruition during the first implementation of the WWW. In particular, in the original proposal it is envisioned "the creation of new links and new material by readers" so that the information's "authorship becomes universal". And also the "automatic notification of a reader when new material of interest to him/her has become available". This became possible only later with the advent of the so called Web 2.0 (Section 2.1.1.1) and the Semantic Web (Section 2.1.2). These aspects, included in the original WWW vision, could be realised only with the recent developments of the Web, and in particular thanks to:

[40] http://www.w3.org/2004/02/skos/ (accessed January 2014)
[41] http://ogp.me/ (accessed January 2014)
[42] http://www.google.com/insidesearch/features/search/knowledge.html (accessed January 2014)
[43] https://www.cia.gov/library/publications/the-world-factbook/ (accessed January 2014)
[44] http://www.freebase.com/ (accessed January 2014)
[45] http://schema.org (accessed January 2014)
[46] http://schema.rdfs.org (accessed January 2014)

- the potential of data and knowledge representation on the Web, which allows to semantically relate and describe any resource (not only documents but also people, concepts, etc.);
- the social and collaborative features offered by Web 2.0 technologies, where users can actively contribute to the Web content and interact with each other increasing the rate of information sharing/production.

While it is quite common to view the Web 2.0 and the Semantic Web as mutually exclusive and competing paths to the Web of the future, the two approaches are in fact complementary. Both face challenges the other can solve, such as how to integrate Web 2.0 data on a Web scale, and how to enable users to create semantically rich annotations. Web 2.0 provides several applications producing and reusing user-generated content, supporting social and collaborative interaction on the Web, and providing engaging user interactions. The Semantic Web vision relies on data published in machine-readable formats, given formal semantics through the use of shared ontologies, and interlinked on a Web scale. By making Web data more open to processing by machines, the Semantic Web fundamentally aims to bring tangible benefits to users.

Starting from Web 2.0 applications producing vast amounts of user-generated content, such as wiki entries, tagged photos, and links joining people in social network, the Semantic Web offers a platform on which publishing data in RDF lowers the barriers to its reuse by others. However, information overload became quickly one of the most prominent concerns on the current Web. The growing volume of data online makes it difficult to understand and to get a comprehensive view of our knowledge.

The idea of the *Social Semantic Web* is that we can organize the world's knowledge while using social media, by leveraging Semantic Web technologies to create synergy between human-readable and machine-understandable data. The Social Semantic Web has its basis on the World Wide Web standards, the added semantic structure of the Semantic Web, and the social connectivity of the Social Web, aiming at bringing the Web to its full potential [Breslin and Decker, 2007] [Breslin et al., 2009] [Shakya, 2009].

Tom Gruber describes his vision of the Social Semantic Web as a move from the *collected intelligence* of the Web 2.0 to a *collective intelligence* [Gruber, 2007]. Semantic Web technologies can "enable data sharing and computation across independent, heterogeneous Social Web applications. By combining structured and unstructured data, drawn from many sites across the Internet, Semantic Web technology could provide a substrate for the discovery of new knowledge that is not contained in any one source,

and the solution of problems that were not anticipated by the creators of individual web sites" [Gruber, 2007]. Such aggregation and filtering would not require significant additional effort by end-users, instead, technologies on the Web should allow for lightweight curation together with the existing social conversations.

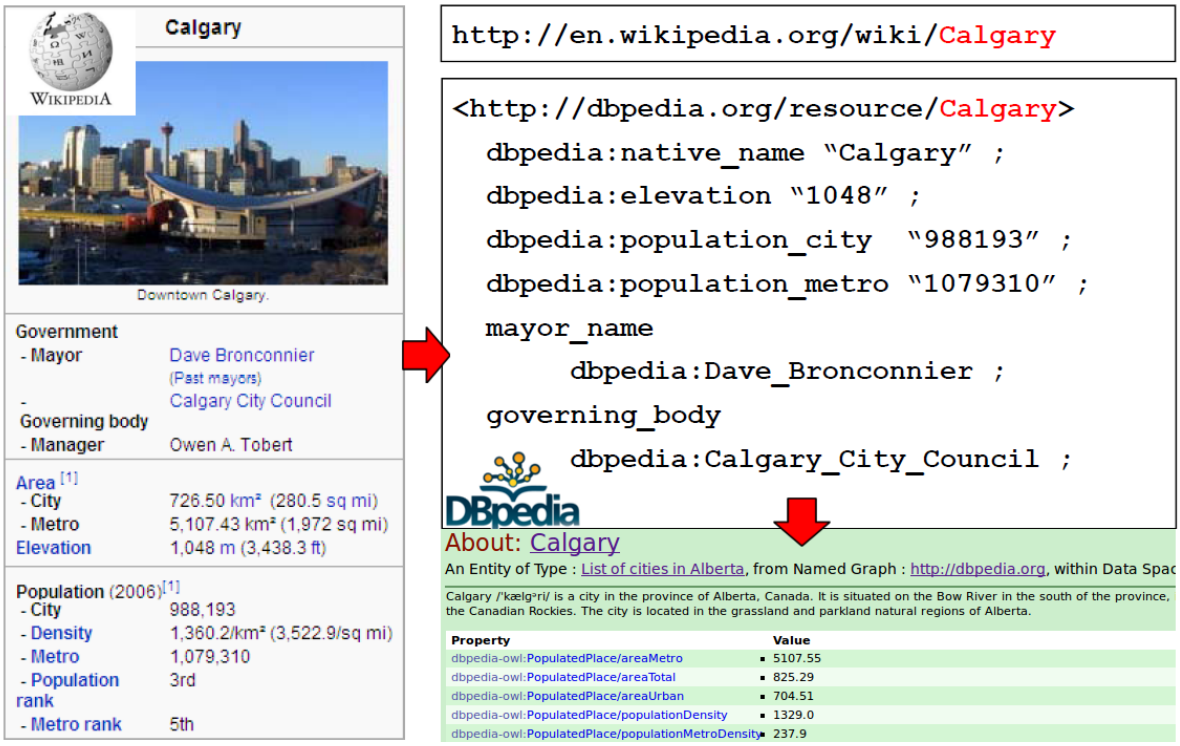

Figure 2.7.: Translation of Wikipedia's structured information into the semantic data of DBpedia[47]

The growth of the Social Semantic Web can be originated from existing media. For instance, two bootstrapping approaches for the Social Semantic Web are: inferring implicit existing structures [Berrueta et al., 2008] and combining ontologies with folksonomies [Specia and Motta, 2007] [Mika, 2007]. By inferring implicit structures, with human analysis of site structures or machine-based data mining, it is possible to lift information from a social website into the Social Semantic Web. Considering for example the Wikipedia case, where templates for Wikipedia articles do not have explicit semantics declared. However, they are already in a semi-structured format that can be automatically translated into semantics (see the DBpedia project [Auer et al., 2007] and Figure 2.7). By combining ontologies with folksonomies, it is possible to improve retrieval and accuracy for the knowledge base while maintaining flexibility during the data entry phase [Passant et al., 2009b].

[47]Image adapted from http://slidesha.re/MvfadP (accessed January 2014)

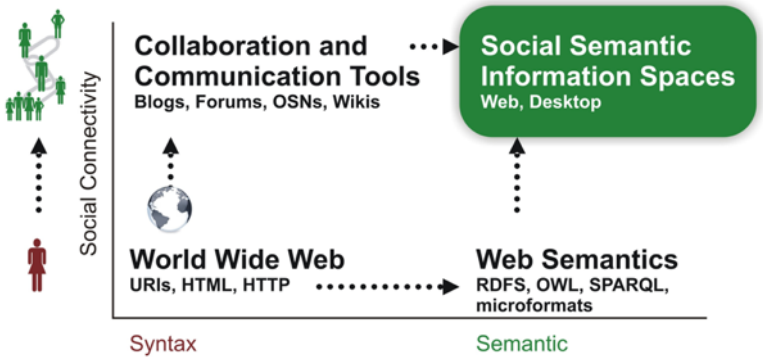

Figure 2.8.: Social Semantic Information Spaces: the convergence between Web 2.0 and Semantic Web [Breslin et al., 2009]

This vision is also shared by Tim Berners-Lee, who described the possibility of having "both Semantic Web technology supporting online communities, but at the same time also online communities can also support Semantic Web data by being the sources of people voluntarily connecting things together" [Berners-Lee, 2005]. This clearly supports the idea of a convergence between Web 2.0 and the Semantic Web (Figure 2.8). A convergence that leads to a Web where the content is provided via social activities and cooperation between end-users, being at the same time machine-processable for autonomous software agents. This is the Web leading to the so called *Social Semantic Information Spaces* [Breslin and Decker, 2006], optimised for both humans and machines; hence, a Web of Data and not only a Web of Documents, where the "desktop" meets the Social Web through the adoption of semantics.

In this thesis we will simply refer to this vision with the term Social Semantic Web: the integration of formal Semantic Web languages, ontologies and schemas on the one hand and Web 2.0 technologies on the other hand.

2.2. Provenance of Data

2.2.1. Definition of Provenance

As a definition of provenance of data we can adopt the W3C Provenance Working Group's definition [Moreau and Missier, 2013a]: "*Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influenc-*

ing, or delivering a piece of data or a thing"[48]. In particular on the Web, provenance can pertain to documents, data, or in general resources over the Web, but also to things in the real world. This is a very pragmatic definition of provenance, especially targeted to the Web context. Another popular definition of provenance in computer science is the following "Provenance as a Process" definition: "*The provenance of a piece of data is the process that led to that piece of data*" [Groth, 2007] [Moreau, 2010].

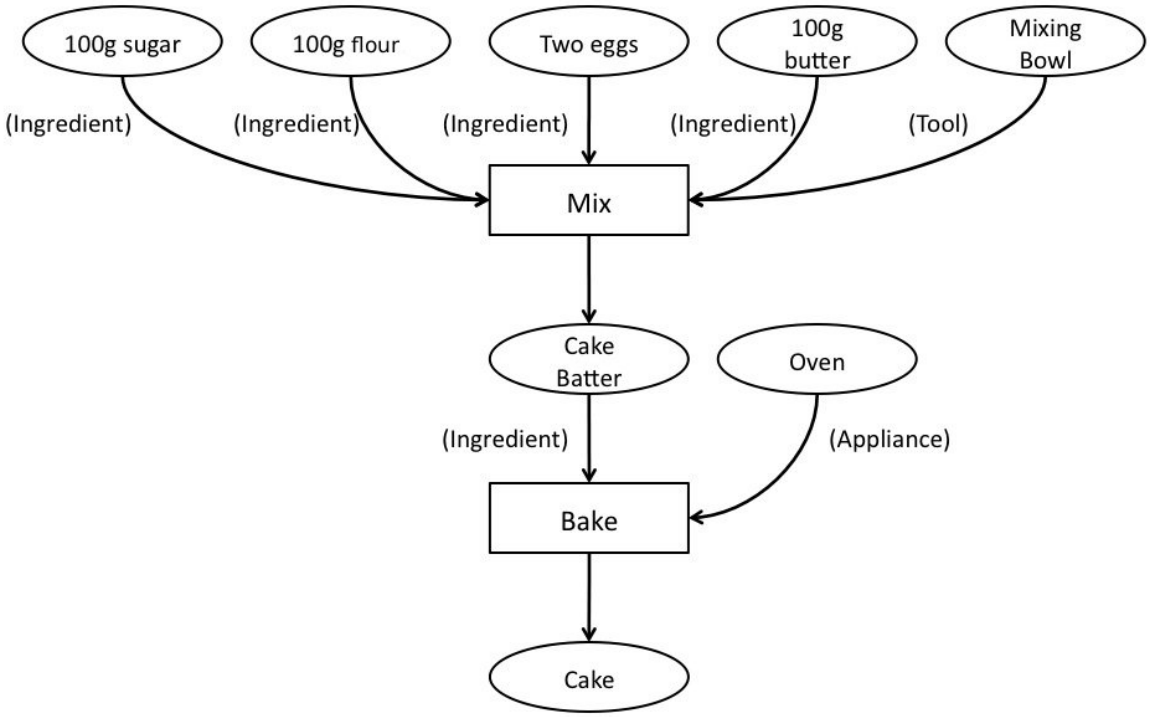

Figure 2.9.: Real world example of provenance records for cake-baking[49]

In fact, quoting the W3C Working Group, "provenance is too broad a term for it to be possible to have one, universal definition - like other related terms such as "process", "accountability", "causality" or "identity", we can argue about their meanings forever (and philosophers have indeed debated concepts such as identity or causality for thousands of years without converging)"[50]. On the Web, provenance is a record (a

[48] http://www.w3.org/TR/2013/REC-prov-dm-20130430/ (accessed January 2014)
[49] From http://tw.rpi.edu/web/project/SPCDIS/Key_Concepts/Provenance (accessed January 2014)
[50] http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/ (accessed January 2014)

form of metadata) that can be created by, exchanged between, and processed by computers. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance [Moreau and Missier, 2013a].

Provenance of data should record the initial sources of information used, as well as any entity and process involved in producing or altering a result or a piece of information. "It offers the means to verify data products, to infer their quality, to analyse the processes that led to them, and to decide whether they can be trusted" [Moreau, 2010]. For example, by using provenance information it is possible to: enable reproducibility of scientific results [Gil et al., 2007] [Davidson and Freire, 2008], or track the authors of particular statements in curated databases [Orlandi and Passant, 2011], or enable reasoning algorithms to make trust assertions about information shared on the Social Web [Carroll et al., 2005] [Artz and Gil, 2007].

2.2.2. Provenance on the Web

The extraction, management and representation of provenance information about data records is not a new research topic. Many studies have been conducted in computer science for representing provenance of data. The majority of work on provenance has been undertaken by the database, workflow and e-science communities. Among all, in [Bose and Frew, 2005] and [Simmhan et al., 2005] the authors provide comprehensive surveys about data provenance management in computer science. The first one provides one of the first surveys in the field applied to a scientific data processing context; while the second one provides a survey and a taxonomy to understand and compare provenance techniques. However, on the Web, we experience a massive and diverse amount of activities for information sharing, discovery, aggregation and filtering. With a growing number of datasets available publicly on the Web, it is important to determine the veracity and quality of these datasets. Hence, it is an additional challenge to identify the original sources and processes producing a particular piece of information on the Web. In this context it is extremely important to track the "lineage" of Web data.

In this regard, a comprehensive survey about provenance on the Web has been published by L. Moreau [Moreau, 2010]. By comparing different models and theories for managing Web data provenance, it is evident the reoccurring presence of three main concepts for modelling data life-cycles: *Actors*, *Processes* and *Artefacts*. Indeed, a mod-

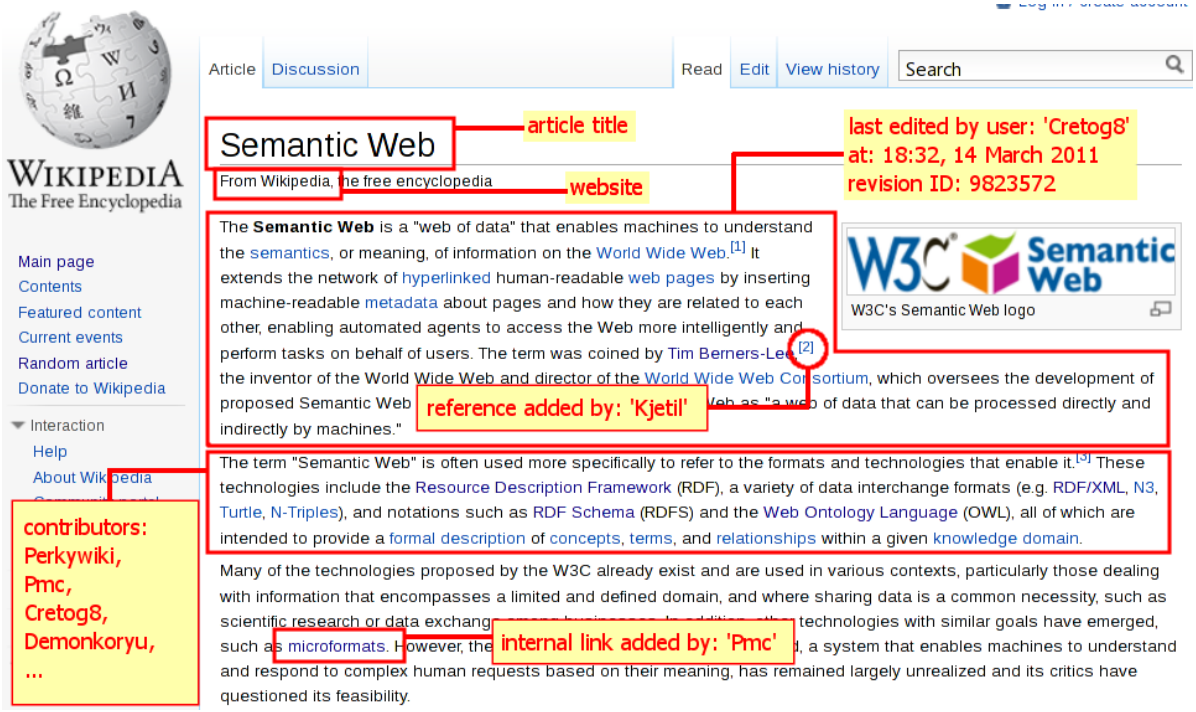

Figure 2.10.: Example of some provenance features that could be extracted from a Wikipedia article. Each article is the result of many changes and actions performed by different users.

elling approach can be "process-oriented", "data-oriented" (the two distinctions made in [Simmhan et al., 2005]), or "actor-oriented" (as proposed in [Harth et al., 2007]). This classification of provenance systems is made on the basis of the subject that is described and its granularity. It can be more suitable to collect provenance about certain types of data products than on others. This decision has to be taken according to the importance of the data or the cost of the provenance collection process. Therefore, it depends whether the focus of the provenance description is more on processes, artefacts or actors. On the Social Web, for instance, it may be particularly appropriate an actor-oriented model [Harth et al., 2007].

In order to standardise provenance systems and their models on the Web, relevant effort has been shown by the W3C with the Provenance Working Group (preceded by the W3C Provenance Incubator Group[51]) The Working Group completed its activity on the 19th of June 2013, publishing W3C Recommendations and documents supporting "the widespread publication and use of provenance information of Web documents, data,

[51] established in September 2009. http://www.w3.org/2005/Incubator/prov/ (accessed January 2014)

and resources"[52]. In particular, the *PROV* Family of Documents [Missier et al., 2013] — including a Data Model (PROV-DM) [Moreau and Missier, 2013b] and an Ontology (PROV-O) [Lebo et al., 2013] — for provenance interchange on the Web has been published as a Recommendation. PROV defines a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or any artefact in the world[53]. As an overview, the key PROV concepts are depicted in Figure 2.11.

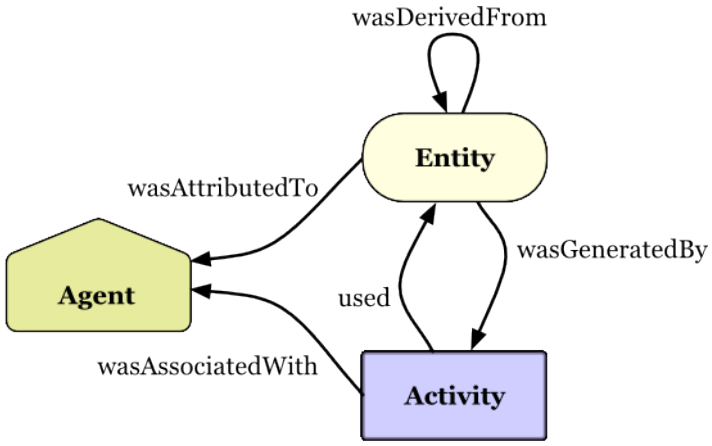

Figure 2.11.: Intuitive overview of PROV, key concepts[54]

In this thesis, we agree with W3C's vision in that providing this information as RDF would make provenance metadata more transparent and interlinked with other sources. It would also offer new scenarios on evaluating trust and data quality on the top of it. Requirements for provenance on the Web, along with several use cases and technical requirements have been provided by the working group. Many additional activities and documents have been included in the final report of the activities of the Incubator Group[55]. We invite the reader to consult this document in order to have more detailed information not only about PROV but also on the requirements for provenance needed in this work. In particular, requirements and use cases, in terms of key dimensions that concern provenance, are summarised in Table 2.2.

The work done in this thesis tackles all the aspects of provenance listed in Table 2.2 with regard to provenance in social media websites and the Web of Data. We aim at

[52] http://www.w3.org/2011/prov/wiki/Main_Page (accessed January 2014)
[53] http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/ (accessed January 2014)
[54] From http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/ (accessed January 2014)
[55] http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/ (accessed January 2014)

Category	Dimension	Description
Content	Object	The artefact that a provenance statement is about.
	Attribution	The sources or entities that contributed to create the artefact in question.
	Process	The activities (or steps) that were carried out to generate or access the artefact at hand.
	Versioning	Records of changes to an artefact over time and what entities and processes were associated with those changes.
	Justification	Documentation recording why and how a particular decision is made.
	Entailment	Explanations showing how facts were derived from other facts.
Management	Publication	Making provenance available on the Web.
	Access	The ability to find the provenance for a particular artefact.
	Dissemination	Defining how provenance should be distributed and its access be controlled.
	Scale	Dealing with large amounts of provenance.
Use	Understanding	How to enable the end user consumption of provenance.
	Interoperability	Combining provenance produced by multiple different systems.
	Comparison	Comparing artefacts through their provenance.
	Accountability	Using provenance to assign credit or blame.
	Trust	Using provenance to make trust judgements.
	Imperfections	Dealing with imperfections in provenance records.
	Debugging	Using provenance to detect bugs or failures of processes.

Table 2.2.: Provenance dimensions: a summary of requirements and use cases for provenance identified by the W3C Working Group

using provenance information for understanding user activities on the Social Web and profiling her interests. However, in this dissertation we do not investigate the dimensions *Trust*, *Imperfections* and *Debugging* of Table 2.2 which we still consider as possible future work. More details about our work in relation to provenance are in Chapter 4.

2.3. User Modelling

During the last decade we have assisted to the growth of Web applications using or collecting data on their users and their behaviour in order to provide adapted and

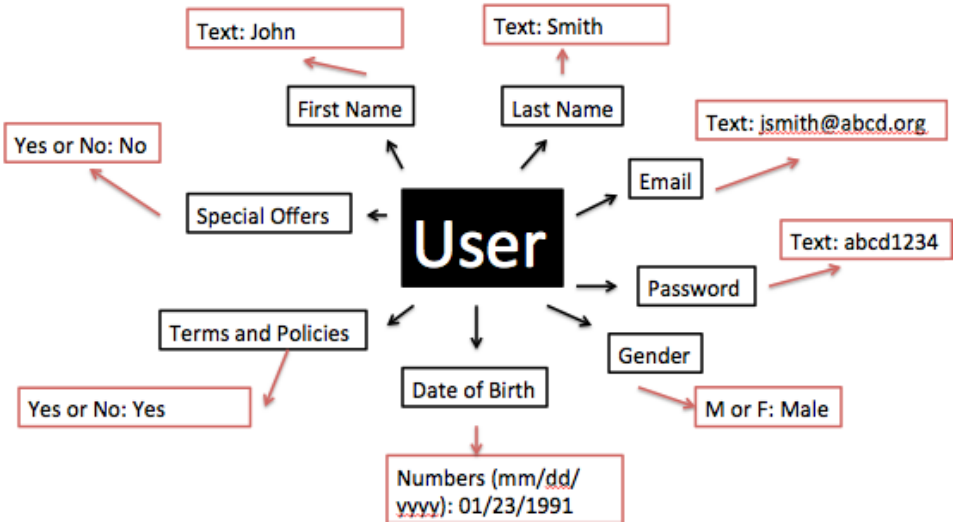

Figure 2.12.: Example of a user model[56]

personalized contents. This caused the need for exchange, reuse, and integration of their data and user models. A new research challenge then emerged, seeking solutions for user modelling and personalization across application boundaries [Viviani et al., 2010] [Carmagnola et al., 2011]. In this section we describe the current state of the art for the research fields of user modelling and personalization. Particular attention will be given on the connection between these fields and the Semantic Web.

2.3.1. Introduction to User Modelling

User modelling techniques are applied by adaptive Web systems to represent with formal models the interests, knowledge and goals of their users (Figure 2.12). These user models are then necessary to provide a personalized experience for different users, for instance by filtering the content relevant to the user on a website, rearranging elements on a page, or recommending users with similar interests. Approaches for personalization cannot be applied without an accurate understanding of the user. The field of user modelling [De Bra et al., 2010] [Kobsa, 1991] is focused on techniques for the description of user knowledge into user models which constitute the basis for adaptive systems.

According to Brusilovsky et al. [Brusilovsky and Henze, 2007], Web personalization now constitutes a large research field that includes communities such as Web science,

[56]Image by Megan Rawley (CC-BY-SA-3.0) http://commons.wikimedia.org/wiki/File%3ACody_User_Model.png (accessed January 2014)

hypertext, user modelling, machine learning, information retrieval, intelligent tutoring systems, cognitive science, Web-based education, etc. Our main focus here is in particular on personalization techniques for adaptive Web systems. These techniques can be grouped in three areas [Brusilovsky et al., 2007]: personalization of information retrieval, personalization of browsing and personalization through filtering and recommendation. Web systems for adaptive recommendation are a more specific type of adaptive Web systems which attempt to deduce the users goals and interests from her browsing activity and recommend a list of related content and relevant links to the user [Brusilovsky et al., 2007]. The core of this field is represented by the user modelling research area. Personalization and adaptation are based on complex information related to user's knowledge, activities, interests, social relations, etc. that could be modelled using structured representations and user modelling techniques. Generalizing, this information about a user is typically stored and represented with a *user profile*.

As regards the user modelling field, at the moment we identify three main *challenges*.

- The first one is about how to *retrieve* information about user interests, knowledge, behaviour and social context. In other words the challenge is to find ways to collect all the useful information needed to build user profiles (Section 2.3.2).
- The second important aspect is related to how to *manage* and *represent* user models in an *interoperable* and scalable way. Hence, the goal is the aggregation and exchange of user models between heterogeneous applications, and the accurate representation of complete and global user profiles (Section 2.3.3).
- The third challenge regards the use of *Semantic Web* technologies and the Web of Data in order to enrich *user models* and provide an interoperable and more accurate representation of user profiles (Section 2.3.4).

2.3.2. User Information Retrieval

With regard to the work done on user information retrieval, latest techniques to track the user behaviour have to cope with the current highly dynamic and socially interactive Web applications and have to be extended to collect fine-grained data from user interactions to provide better information for adaptive systems. Additionally, the collected data must be managed in ontologies to share user behaviour information with other adaptive systems. Zhou et al. [Zhou et al., 2005] focus on mining client-side access logs of a single user or client and then incorporate fuzzy logic to generate a usage ontology.

Schmidt et al. [Schmidt et al., 2007] embed concepts into a portal which provides the context for JavaScript events, which are collected and used to adjust the portal. All the relevant user interface elements are linked to a concept ontology containing semantic information about the elements. None of these approaches make full use of semantic technologies. First steps in the direction of semantic technologies are done but they still cannot be applied across heterogeneous applications and lack the necessary extensibility and dynamism (see Section 2.3.4).

Szomszor et al. [Szomszor et al., 2008] investigate the idea of merging users' distributed tag clouds to build richer profile ontologies of interests, using the FOAF vocabulary and matching concepts to Wikipedia categories. The authors experimented with over 1,300 users who showed high activities in both of the two websites Del.icio.us[57] and Flickr[58]. For each user, data about each of the two tag clouds has been retrieved and then merged. The results described in the paper show that, on average, 15 new concepts of interest were learnt for each user when expanding tag analysis to their tag cloud in the other folksonomy. In this case, the user profiles generated are represented using popular lightweight vocabularies such as FOAF.

In this context, very relevant is also the *user identification* aspect. In order for applications to share information about users, mechanisms for the identification of users are necessary. Identity-based protocols such as OpenID[59] or WebID[60] can be used for users to link their different identities on the Web. Google Friend Connect[61] provides an API which exemplifies the use of OpenID and OAuth to integrate registered users, existing login systems, and existing data with new social data and activities. It is based on open standards (OpenID, OAuth and Open Social[62]) and allows users to control and share their data with different social websites. Moreover, the WebFinger[63] protocol documents a way to get a XML file describing how to find a user's public metadata from that user's email address like identifier, providing then information about existing user accounts linked to that user.

[57] http://www.delicious.com/ (accessed January 2014)
[58] http://www.flickr.com (accessed January 2014)
[59] http://openid.net (accessed January 2014)
[60] http://www.w3.org/wiki/WebID (accessed January 2014)
[61] http://code.google.com/apis/friendconnect/ (accessed January 2014)
[62] http://code.google.com/apis/opensocial/ (accessed January 2014)
[63] http://code.google.com/p/webfinger/ (accessed January 2014)

2.3.3. Architectures for User Model Interoperability

Latest developments on user modelling involve interoperability and portability of user models [Carmagnola et al., 2011] [Viviani et al., 2010]. The rapid growth of user adaptive and social systems, collecting information about users, led to the replication of user data over many applications. This inevitably conducted researchers to deal with an important challenge: user model interoperability. In other words, the process of exchanging distributed and heterogeneous user data across applications [Vassileva, 2001]. User model interoperability would provide several advantages under quantitative and qualitative aspects. It allows, the collection of more data and more accurate data about users; the acquisition of user modelling functionalities that systems do not themselves implement; a solution to the well-known "cold start" problem (Section 2.4) during the user model initialization phase and a consequent speed-up of this phase. On the other hand achieving interoperability on the Web, a completely open and dynamic environment, is a complex and challenging task that requires open and agreed standards and a high level of alignment of the involved systems.

In the context of user model representation and management increasing relevance is attributed to the interoperability of the representations. Applications typically store their user information in a proprietary format. This leads to a distributed Web model of a user with several partial user models in different applications potentially duplicating information. Therefore, the challenge is to solve the heterogeneity of the user models. Current research on user model management and aggregation emphasizes two different strategies [Kuflik, 2008].

- The first strategy introduced in [Berkovsky et al., 2008] uses a generic *user model mediation* framework with the goal of improving the quality of recommendations. The actual UM mediation in the framework is done by specialized mediator components which translate the data between different models using inference and reasoning mechanisms. In their subsequent work [Berkovsky et al., 2009] Berkovsky et al. still focus on cross-representation mediation of user models describing its practical implementation and evaluating the outcome of the collaborative to content-based filtering user model mediation. As they state in their paper "the mediation procedure allows bootstrapping the empty UMs and enriching the existing UMs in a content-based recommender system, and, as a result, more accurate recommendations are generated".

- The second strategy focuses on the *standardization of user models* to allow data sharing between applications. Heckmann [Heckmann et al., 2005c] proposes an ontological approach, the General User Model Ontology (GUMO), as a top level ontology for user models and suggest the ontology to be the standard model for user modelling tasks. Another standardization approach is to define a *centralized user modelling system* that is used and updated by all connected applications [Korth and Plumbaum, 2007].

The drawback of the mediation layer approach is the effort needed to aggregate such heterogeneous user models, while standardized user models suffer from the lack of a common standard.

Different solutions and architectures have been proposed in order to solve the interoperability problem. They can be categorized in three types of approaches: *centralized*, *decentralized* and *mixed*. This categorization is mainly based on two factors: the *physical storage*, or where the user data is maintained, and the *conceptualization* of the model, that is "how the user model component is conceived in terms of being shared or not between systems" [Carmagnola et al., 2011]. A *centralized* approach then represents systems that are both physically and conceptually centralized; *decentralized* approaches are physically and conceptually distributed; *mixed* approaches refer to systems that are physically decentralized and conceptually centralized. In most of the cases standardization-based approaches are conceptually centralized and mediation-based ones are conceptually decentralized. To note that so far user modelling systems are evolving from centralized to decentralized architectures. This tendency is motivated mainly by the difficulties in developing and adopting a unique and common user modelling standard and by the intrinsic decentralized nature of the Web. Moreover centralized systems are by definition affected by the single point of failure problem and by the privacy and security of users' information which is all stored in a single point. A comprehensive list of user modelling systems appropriately categorized is provided by two recently published surveys by Carmagnola et al. [Carmagnola et al., 2011] and Viviani et al. [Viviani et al., 2010]. We refer to these two publications for a complete overview of the state of the art in this research field. In the following subsection we select and describe only the work that is particularly relevant to our dissertation.

2.3.3.1. Review of User Model Interoperability Systems

Centralised Systems Latest developments and examples of centralized architectures are described in PersonisAD [Assad et al., 2007], UMS (User Modelling Server) [Kobsa and Fink, 2006], MUMS (Massive User Modelling System) [Brooks et al., 2004].

Assad et al. [Assad et al., 2007] developed a framework called *PersonisAD* that aims at supporting the development of context-aware applications using distributed user models. The framework is targeted at ubiquitous applications and supports the management of different kinds of models, such as models of users, places, sensors, services and devices. Therefore, not only user data is exchanged, but also data about the environment. However, an application complying with this framework has to use a common user model at an ontological level of the components in the environment, in order to have knowledge about the components themselves and about the different contexts in which the user models are organized.

Kobsa and Fink [Kobsa and Fink, 2006] (see also [Kobsa, 2007] and [Fink, 2003]) developed a *User Modelling Server (UMS)* based on the *Lightweight Directory Access Protocol (LDAP)*. It allows external applications to submit and retrieve information about users whose models are represented in the system. Therefore, it provides a user modelling service to other applications and is capable of representing different types of models, from user profiles to system and service models. The type of exchanged data is strictly related to users (demographic data, interests and preferences) and application usage.

Brooks et al. [Brooks et al., 2004] in their work describe the *MUMS* system, a *Massive User Modelling System*. It is a centralized system that provides a user modelling/adaptation service, it supports "the just-in-time *production, delivery* and *storage* of user modelling information". It is suitable for describing any domain that can be expressed in RDF/OWL. Hence, it uses Semantic Web techniques and standards. In order to represent the users it adopts a shared user model ontology and all the managed information is expressed in RDF. The interaction between the user data producers and user modellers systems utilizing the data is mediated by a central broker component, while the architecture and the communication layer is Web service based.

Decentralised Systems As regard decentralized approaches, in [Mehta et al., 2005] Metha et al. propose a standardization-based approach using a common ontology-based

user context model (UUCM — Unified User Context Model) as a basis for the exchange of user profiles between multiple systems. Cross system personalization is then obtained relying on an unified profile for each user which is stored inside a "Context Passport" [Niederée et al., 2004]. Further developments of this work by Metha et al. are described in [Mehta and Nejdl, 2007] where the authors propose machine learning techniques for automatically matching user models. Dependencies between profiles are computed analysing data provided by users sharing their profile across different systems and learning from that population. The UUCM is also encoded as an RDF Schema augmented with OWL expressions enabling exchange possibilities with other Semantic Web enabled systems.

Another example of decentralized architectures for user modelling is presented by Heckmann et al. [Heckmann et al., 2005a] where user-adaptive systems exchange user information using UserML [Heckmann, 2003], a RDF-based user model exchange language, and the *General User Model Ontology* (GUMO) [Heckmann et al., 2005b], an ontology for the uniform interpretation of decentralized user models. This is another example of a standardization-based approach as the GUMO ontology is proposed as the uniform interpretation of distributed user models in Semantic Web environments. It is so far the most comprehensive user modelling ontology but at the same time it is very extensive and it might be complex to implement in a real system. Moreover this vocabulary has to be adopted by the systems that want to exchange user models, so an *a priori* agreement between the systems is necessary, in the same way as in [Mehta et al., 2005] previously described.

In [Carmagnola and Dimitrova, 2008] [Carmagnola, 2009] a new approach for user model interoperability is proposed. The authors propose a framework that "deals with semantic heterogeneity of user models and automates the user model exchange across applications". It is inspired by Semantic Web technologies and represents an intermediate solution which combines both a flexible user model representation and an automatic semantic mapping of user data across different systems. An algorithm based on evidential reasoning has beed developed in order to create mappings between concepts and values present in different user models and measure their similarity (*Object Similarity Algorithm* and *Property Similarity Algorithm*). User models are represented and exchanged in RDF and queried using SeRQL (Sesame RDF Query Language)[64].

[64] http://www.openrdf.org/doc/sesame/users/ch06.html (accessed January 2014)

-	Architecture	Pros	Cons
Assad et al. 2007 *(PersonisAD)*	Centralized; Standard-based;	User + environment models; Scrutable user models;	No Semantics; Common user model;
Kobsa & Fink 2006 *(UMS)*	Centralized; Mediation-based;	No common user model;	No Semantics; Based on LDAP;
Brooks et al. 2004 *(MUMS)*	Centralized; Standard-based;	Real-time service; User models in RDF/OWL;	Common user model;
Metha et al. 2005	Decentralized; Standard-based;	Machine learning for model matching; UUCM user models in RDF/OWL; User + environment models;	Common user model;
Heckmann et al. 2005	Decentralized; Standard-based;	GUMO ontology in OWL; User + environment models; Scrutable user models;	Common user model; Complexity of GUMO ontology;
Carmagnola et al. 2009	Decentralized; Mediation-based;	User model in RDF; Reasoning for user model mapping;	No scrutable user model;

Table 2.3.: Comparison of the reviewed systems targeting user model interoperability

Comparison In this section we show a comparison table including the systems for user model interoperability that we reviewed previously in Section 2.3.3. In Table 2.3 we display only the systems with a complete implementation: from the information retrieval task, to the mapping of user concepts and values, to the provision of integrated user profiles or a personalization service available to other external applications. Moreover, this is not a complete table including all the applications in the state of the art, but it represents a selection of some of the most interesting systems from our perspective considering our research goals. In Table 2.3 we categorise the systems according to their architecture and then we list the positive and negative aspects that we see in those implementations. Some of these aspects are subjective and somehow influenced by our research background. For further details please refer to the description of the systems in the previous subsections.

2.3.4. Semantic Web Technologies for User Modelling

Interesting research that bridges the gaps between user information retrieval/profiling and the Semantic Web has been presented by Szomszor et al. [Szomszor et al., 2008]. The authors investigate the idea of merging users' distributed tag clouds to build richer profile ontologies of interests, using the FOAF vocabulary and matching concepts to Wikipedia categories. We previously described this work in Section 2.3.2 and it is particularly relevant that the authors demonstrate the benefits of the amalgamation of multiple Web2.0 user-tagging histories in building personal semantically-enriched profiles of interest. The user profiles generated are also represented using a popular lightweight vocabulary such as FOAF.

A survey on adaptive systems adopting Semantic Web technologies is provided in [Torre, 2009]. The author describes a classification of adaptive systems based on a distinction between *strong semantic techniques* and *weak semantic techniques*. The former regards systems based on the Semantic Web approach and the latter regards technologies that basically aim at annotating resources in order to enrich their meaning. The survey is mainly focused on weak semantic approaches, these are particularly successful in contributing to user modelling tasks especially when combined with social tagging features. On the other hand strong semantic techniques are more suitable for user knowledge integration and reasoning. The authors also suggest that a category of mixed approaches is growing and it benefits of the advantages of both the technologies in different tasks. The analyzed tasks belongs to the topics of domain modelling and management, context modelling and management, adaptation, personalization and privacy. The authors provide a matrix summarizing the reviewed systems on the basis of the semantic technology that was used and the task it was used for.

Relevant related work on Semantic Web applied to user modelling and personalization has been done by Aroyo et al. [Aroyo and Houben, 2010]. In this work the authors highlight the challenges they see in the near future for user modelling and the adaptive Semantic Web. Furthermore, a review of the research in this field is provided. In the state of the art review the authors analyse the differences between past user modelling solutions (in traditional "*closed*" Web-based or application-based systems) and new research on "*open*" and Semantic Web based solutions. The fundamental tasks identified by the authors that contribute to user modelling are: user identification, user property representation, and sharing adaptation functionalities. An analysis of some of the possible solutions to these tasks is provided by the authors, and relevant related work is

also presented. Moreover the authors provide a set of challenges on this research field describing possible future developments and scientific questions.

The major question in user identification investigates how to identify a person on the Web, her multiple identities across different applications and what are the trust and privacy aspects involved. As regards user knowledge the main challenge is to find ways to share user models, and this implies the definition of common vocabularies and interoperable representations of objects and values of user properties. Finally in their work Aroyo et al. highlight an important aspect about the openness of the Web of Data and the related implications of this on users' experience: an open approach to user knowledge would produce different new use cases and knowledge management approaches, especially users should then be able to inspect and edit their own data. Related and more practical work by the same authors and others is described in [Schopman et al., 2010] and [Van Aart et al., 2009] where, as part of the NoTube project, by using the Linked Data cloud, semantics can be exploited to find complex relations between the user's interests and background information of TV programmes, resulting in potentially interesting recommendations. In another paper [Denaux et al., 2005] Denaux et al. present how interactive user modelling and adaptive content management on the Semantic Web can be integrated in a learning domain to deal with common adaptation problems (e.g. cold start, inaccuracy of assumptions, knowledge dynamics, etc.).

Finally in the previous section we already described the work done by Carmagnola et al. [Carmagnola and Dimitrova, 2008] [Carmagnola, 2009] representing one of the most advanced user modelling systems adopting semantic technologies. The use of RDF for representing user models and the reasoning capabilities implemented with a "SPARQL-like" language (SeRQL) on top of the user models in order to obtain automatic mapping between heterogeneous concepts are the strongest points of their implementation. A drawback of their system is the lack of *scrutable* user models, it is not possible for a system user to consult her user model created by the application.

As we described previously, some of the systems for user model interoperability analysed use RDF or OWL to represent user models however the user models created cannot be shared or integrated easily with other different systems or on the Web of Data because of the complexity and particularity of the ontologies used. Moreover, in almost all the cases, reasoning capabilities on top of the user data are not implemented using Semantic Web technologies.

2.4. Personalisation

From a marketing perspective, personalisation is "the adaptation of products and services by the producer for the consumer using information that has been inferred from the consumer's behavior or transactions" [Montgomery and Smith, 2009]. On the Web, this generic definition is still valid, as Web users can be seen as consumers of content which is made available by some publishers or producers via Web documents or Web services. In this regard, even though it is difficult to provide a unique definition of personalisation, we agree with the following one retrieved from Wikipedia[65]: "Personalization technology enables the dynamic insertion, customization or suggestion of content in any format that is relevant to the individual user, based on the users implicit behaviour and preferences, and explicitly given details". It is important to highlight two main elements of this definition: the first one is that the personalised content on the Web can be either suggested (i.e. by a recommender system [Montaner et al., 2003]) or just dynamically customised (e.g. adaptive hypermedia [Brusilovsky, 2001]); the second factor is that the information about the user can be either automatically inferred from her activities or explicitly provided by the user to the personalisation system. Indeed, we distinguish three main personalisation methods: **Implicit**, **Explicit** and **Hybrid**; according to the way user information is collected. User information is collected and modelled in most personalisation systems according to predefined *user models* into personal *user profiles*.

The most popular personalisation systems on the Web so far are recommender systems (Figure 2.13). Especially in the last decade the Web experienced a steep increase in the number of Web pages and services providing its visitors suggestions and recommendations about products, topics, users, etc. Personalised recommendations have demonstrated their effectiveness in enhancing users' experience of searching, exploring and finding new and interesting content [Heitmann et al., 2012b] [Montaner et al., 2003]. Especially in the context of e-commerce and Social Web recommendations of users and interests. The typical recommender system is divided into three main components [Burke, 2002]:

- **background data**, representing the knowledge base that the system has about the objects to be recommended;

[65] And originally from: Doman, James. "What is the definition of "personalization"?". Quora. Retrieved 19 March 2012.

[66] Image from http://e-strategyblog.com/2011/06/daily-numbers-dear-john (accessed January 2014)

Figure 2.13.: Personalised recommendations offered to a user on the Amazon e-commerce website[66]

- **input data** or **user model**, representing the user information provided in order to make recommendations;
- **recommendation algorithm**, which combines background and input data according to different strategies in order to provide recommendations.

According to Burke's classification of recommender systems [Burke, 2002] there are four main groups of algorithms: (i) *collaborative filtering*, (ii) *content-based*, (iii) *knowledge-based* and (iv) *hybrid*. Moreover, an additional group of recommendation algorithms is considered in [Heitmann et al., 2012b]: (v) *graph-based* recommender systems. In the following subsections we briefly summarise the different recommendation algorithms following the same categorisation described in [Heitmann et al., 2012b] and [Burke, 2002].

Collaborative Filtering This method aims at predicting the interests of a user by collecting preferences from other users [Herlocker et al., 2004]. The assumption is that if a user X has the same opinion as a user Y on a particular issue, X is more likely to have Y's opinion on a different issue than to have the opinion of another random user on that different issue. Hence, similarity measures between users are used to make

the recommendations. The input data consists of a user profile with ratings for one or more items, the background data is the set of all the other ratings of the other users. Similarity measures are computed for the input data against the background data in order to recommend items or users [Sarwar et al., 2001]. A popular example of collaborative filtering is the item-to-item collaborative filtering algorithm adopted by Amazon[67], also known as "*people who buy x also buy y*".

Content-Based Recommendation The background data for this group of recommender systems is a set of items with related descriptive features (content features related to the items such as metadata, textual description, links, tags, timestamps, etc.). The input data is a user profile containing a user's description of preferences through similar content features as in the the background data. Content features of background data and input data are matched with similarity measures for the recommendations [Pazzani and Billsus, 2007]. Popular examples of a content-based recommender systems are online music radios such as Pandora[68] or movie recommenders such as IMDB[69]. Pandora, for example, offers an online radio service that plays music with similar characteristics to that of a song provided by the user as an initial seed.

Knowledge-Based Recommendation Knowledge-based systems are similar to the content-based ones, with the difference that usually the knowledge base (both background data and input data) contains explicit functional knowledge about how certain item features meet user needs [Burke, 2007]. Hence, an algorithm can reason about the relationship between a need and a possible recommendation.

Graph-Based Recommendation This category of recommenders, as nicely summarised in [Heitmann et al., 2012b], aims at exploiting the *social graph* made of online social interactions of users and/or their *interest graph* built with the users' interests and their connections. Specific graph based algorithms are designed to traverse and analyse these graphs for the recommendations. In particular, so far two types of algorithms have been employed: Semantic Distance [Passant, 2010] and Spreading Activation [Heitmann, 2012] [Marie et al., 2013].

[67] http://www.amazon.com (accessed January 2014)
[68] http://www.pandora.com (accessed January 2014)
[69] http://www.imdb.com (accessed January 2014)

Hybrid Algorithms A hybrid approach combines multiple techniques together to achieve some synergy between them. Many possible options are in this category of systems, we refer to [Burke, 2007] for more details on the topic. Netflix[70], the provider of on-demand Internet streaming media, is a popular example of hybrid recommender system. They combine collaborative filtering and content-based filtering as they recommend movies by comparing the watching habits of similar users as well as by analysing the characteristics of the films that a user has watched and rated.

Common challenges for recommender systems — and personalisation systems in general — are:

- **Cold Start**: this problem verifies when a system requires an initial large amount of data about its users in order to provide accurate recommendations.
- **Sparsity**: this happens when the personalisation system has a large amount of items and data in its background data that needs to be rated by the users of the systems so that the algorithm can provide accurate suggestions. When not enough item ratings and users belong to the system we have the so called "sparsity" problem.
- **Scalability**: often a large amount of computational power is necessary to calculate recommendations, when a large number of items, descriptions and users are in the system.

In this section we summarised the main concepts related to personalisation systems. This section is not intended as a complete reference on the topic, as the main focus of our work is not on the personalisation systems themselves but on user profile data. As we have just described, this data (being it either background or input data) is crucial for these systems in order to provide complete and accurate personalisation. Moreover, the terminology and definitions introduced here will be used throughout the thesis.

[70] http://www.netflix.com (accessed January 2014)

Chapter 3

Characterisation of Social Media and Aggregation of Social Web Data

3.1. Introduction

The main contributions of this thesis focus on methods for the retrieval and aggregation of user interests using Social Web data, distributed across heterogeneous social media sites. Therefore, an initial characterisation of the current different types of social media is fundamental. The basis of our approach lies on top of these social platforms and their data. The idea behind the methodology presented in this dissertation is that concepts and objects of interest of Social Web users can be extracted by analysing content and activities produced and performed on social media. Users on the Social Web interact with each other, create/share content and express their interests on different social websites with many user accounts and different purposes (Figure 3.1). Activities such as posting a personal status message, commenting a media object, publishing a blog post, liking a friend's post, or editing a wiki article, are just a portion of the diverse types of actions available to current Web users. The content produced and the activities performed on social media express, implicitly or explicitly, different kinds of user preferences.

On each social website personal information, consisting of a portion of the complete profile of the user, is recorded. With respect to "complete user profile" we intend the full set of personal information belonging to a person obtained by aggregating the distributed partial user profiles on each Social Web system. Each partial user profile might

[1]Image from http://slidesha.re/1fPdO8N (accessed January 2014)

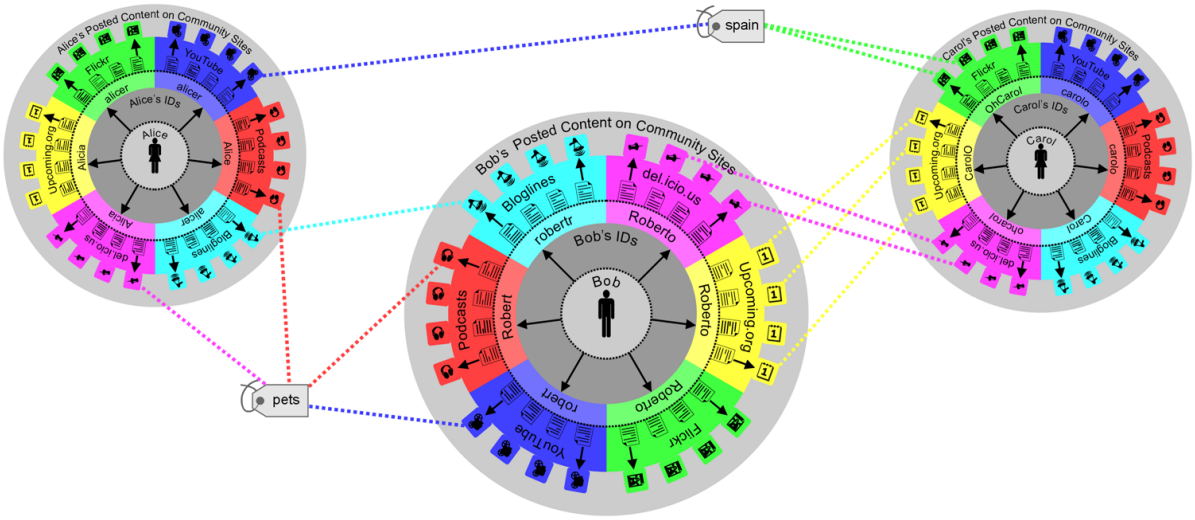

Figure 3.1.: Users on the Social Web create and consume content using different user accounts. They interact within various communities and share content and interests. An illustration by Breslin et al.[1]

contain the user's personal and contact information, her interests, activities and social network of contacts. In this thesis we focus on user profiles of interests as structured and ranked collections of concepts relevant to the users. These details are typically used by applications for personalisation and recommendation purposes. All the distributed user profiles on the Web represent different *facets* of the user therefore their aggregation provides a more comprehensive picture of a person's profile [Abel et al., 2010a]. Aggregation of user profiles brings several advantages: it allows for information reuse across different systems, it solves the well-known "cold start" problem of personalisation systems (Section 2.4), and provides more complete information to each individual Social Web service. However, the aggregation process is a non-trivial problem which derives from the most popular data integration issues: entity matching, duplicates/conflicts resolution, heterogeneity of the sources' data models — and the consequent need of a common target data model — being the most important ones.

Using standard semantic technologies to represent the data sources would help in solving these issues and it would provide a unified representation of the target data model. Furthermore, a complete semantic representation and management of the provenance of user data addresses the duplicate/conflict resolution issues, since it would allow to track the origins of the data at any point of the integration process [Hartig and Zhao, 2010]. Several approaches for aggregating and representing multi-domain user models have been presented in the state of the art so far (see Section 2.3.3) but in most of the cases they are not aimed at defining a standard, source-independent, architecture that allows for

interoperability and integration of profiles of interest on the Web of Data. The use of the best Linked Data principles and the integration with the Web of Data is crucial, as it automatically provides a standard "platform" for the representation of the user data with popular vocabularies. It also enables for semantic data enrichment using the many open datasets on the LOD (Linked Open Data) cloud. At the same time approaches that aim at integrating user models with the Web of Data [Szomszor et al., 2008] are system dependent and do not focus on aggregation of user data from different sources.

The current chapter will provide a characterisation of the various existing types of social media, describing their distinctive features and mentioning some of the most popular examples existing on the Web (Section 3.2). Moreover, the most important vocabularies and ontologies for the representation of social media sites (their structure, objects, actors and interactions) will be outlined (Section 3.3.1). Finally, we will conclude the chapter with a description of a methodology for interlinking online communities and aggregating social data from multiple sources (Section 3.3). The validity of this methodology is evaluated through its implementation on a particular use case scenario. In particular, an application enabling search on heterogeneous wiki systems will be described (Section 3.4).

3.2. A Characterisation of Social Media Systems

3.2.1. Social Media: Definition, Features and Evolution

Social media has been popularly defined as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content" [Kaplan and Haenlein, 2010]. As previously described in Section 2.1.1, the Web 2.0 led to a second generation of Internet-based services such as blogs, wikis, social networking services, etc. These services provide users with different ways and tools for the creation and exchange of content combining many types of digital media. Not only text but also photos, videos and audio files can be edited and shared in different ways. From this important and distinctive feature derives the name "social media", attributed to this entire group of Web applications.

Many categorisations of social media have been proposed according to: different features offered to the users, particular structure of the website, interactions between the users and the shared objects. In particular, Kaplan et al. distinguish the main types of

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

Figure 3.2.: Classification of social media by Kaplan et al., from [Kaplan and Haenlein, 2010]

social media using sociological concepts such as "*self-presentation*" and "*self-disclosure*" combined with "*social presence*" and "*media richness*" [Kaplan and Haenlein, 2010]. The result is a classification scheme composed of six types of social media, as depicted in Figure 3.2. While this classification obtained wide agreement, especially among sociologists and economists, it is becoming however quite generic and outdated. In fact, since 2010 the boundaries between these different categories have become extremely blurred. New social platforms emerged combining the characteristics of different existing ones or introducing novel social features of interaction. For example, Twitter[2], one of the most popular microblogging services, has shifted towards a richer social networking Web platform [Kwak et al., 2010]. At the same time, services such as Reddit[3] (a social news service), Quora[4] (question-and-answer website), Foursquare[5] (a location-based, mobile, social networking service), etc., are also difficult to categorise.

Other attempts in characterising social media have been proposed in many research fields: from marketing [Kietzmann et al., 2011], to enterprise information systems [Subramaniam et al., 2013], to data mining and knowledge discovery [Barbier et al., 2013]. However, in this thesis we distinguish social media according to their structure and the way users produce, modify and share information. This "pragmatic" view of social media is widely accepted in knowledge representation contexts and by the Social Semantic Web research community [Breslin et al., 2009]. In these communities the aim is to represent and capture the knowledge generated through these popular Social Web tools. This is possible by analysing user interactions, social activities, the generated data and its evolution. Simplifying, this can be generically modelled, as described in Section 2.2.2,

[2] http://twitter.com (accessed January 2014)
[3] http://www.reddit.com (accessed January 2014)
[4] http://www.quora.com (accessed January 2014)
[5] http://www.reddit.com (accessed January 2014)

with an *Agent* - *Activity* - *Entity* model such as the PROV provenance data model (Figure 2.11).

Therefore, to provide an essential characterisation of social media, we follow a straightforward approach, as described in this chapter and the following sections. First, we provide an overview of the main social media platforms, the current prominent tools and communities, by examining the state of the art. Second, we analyse vocabularies and ontologies published and developed by knowledge representation experts and communities of Web developers, companies and enthusiasts. We restrict our focus on vocabularies designed for describing Social Web communities, networks, relationships and media. The fact that these vocabularies are usually constantly updated and designed by a large community of experts, provides us with a characterisation of social media which is both widely agreed and sound.

The following subsections provide an overview of the main social media platforms identified on the state of the art. In Section 3.3.1 we describe the Social Web vocabularies analysed for our purpose.

3.2.2. Social Networking Services

Social networking services are among the most prominent types of social media sites. They enable users to create a personal profile and define any kind of relationship to friends, or to other users in general [Boyd and Ellison, 2008]. They allow people to manage a representation of their social network and make it available to other users. Typically, users communicate via direct messages (either public or private) or public comments on shared media objects or profile pages. Often users are given the ability to create groups or events sharing some common interest or affiliation. Additionally, social network services may include features such as photo-sharing, online games or blogging. The most widely used social network service is Facebook[6] which was launched in 2004 and as of December 31, 2013 had accumulated 1.23 billion monthly active users [Facebook, 2013]. Facebook currently ranks second on the list of most popular websites[7], second only to the Google search engine[8]. Currently, other popular services are for instance Linkedin and Google+[9].

[6] www.facebook.com (accessed January 2014)
[7] According to Alexa (www.alexa.com, accessed January 2014)
[8] www.google.com (accessed January 2014)
[9] Respectively www.linkedin.com and http://plus.google.com (accessed January 2014)

Because of the sensitive nature of the data, social networking sites generally expose less content publicly on the Web than other types of social media. The majority of the social networking platforms offer several privacy options for the user profiles and some enable users to have fine-grained control over what personal content is made publicly available. Additionally, users may join interest user groups, organized by workplace, school or other characteristics, and categorize their friends into lists. In general, because of the personal nature of the communications, users tend to express a wide range of particular and private interests on these networks.

3.2.3. Wikis

The first wiki system was developed by Ward Cunningham in 1994, under the name WikiWikiWeb [Leuf and Cunningham, 2001]. In Hawaiian the word "wikiwiki" means "quick". The original definition of a wiki describes it as "The simplest online database that could possibly work" [Ward and Bo, 2002]. Wikis are web sites that can be collaboratively edited by anyone. Pages are written in a simple syntax so that even novice users can easily edit pages [Wagner, 2004]. The syntax consists of simple tags for creating links to other Wikipages and textual markups such as lists and headings. The user interface of most Wikis consists of two modes: in *reading mode*, the user is presented normal webpages that can contain pictures, links, textual markup, etc. In *editing mode*, the user is presented an editing box displaying the Wiki syntax of the page (containing the text including the markup tags). During editing, the user can request a preview of the page, which is then rendered by the server and returned to the user.

Many Wiki engines exist for anyone who wants to setup a Wiki, most of these engines are open-source. Many sites run a Wiki as a community venue, enabling users to discuss and write on topics. For example, many open-source projects have a documentation Wiki, where users can collaboratively add documentation about the project. The burden of editing is thus shared over the whole community, while still allowing anybody to quickly find relevant documentation (which is harder in e.g. a forum or bulletin board) [Wagner, 2004].

Popular Wikis such as Wikipedia[10] can grow very fast, since interested visitors can edit and create pages at will. Wikipedia is a collaboratively edited, multilingual and free Internet encyclopedia. Wikipedia offers 30 million articles in 287 languages, with over

[10] http://www.wikipedia.org (accessed January 2014)

4.3 million in the English Wikipedia[11]. Almost all of its articles can be edited by anyone having access to the site [Nov, 2007]. It is the largest and most popular general reference work on the Internet having an estimated 365 million readers worldwide[12] [Fallis, 2008].

3.2.4. Blogs

A blog, or a weblog, is a website that contains a set of posts ordered in reverse chronological order. The posts could be of any length and could include links, pictures or other media objects. Blogs are normally maintained by a single author, but sometimes a small group of editors curate the site together. However, every single post is always associated to a single author. Hence, the social and collaborative side of a blog does not reside on its publishing part but rather on the interaction with the readers. Readers of a blog are generally allowed to publicly share their comments on the blog posts or on other readers' comments, therefore creating a community of readers around the blog. The collective community of all blogs is known as the *blogosphere* [Klamma et al., 2007]. Since all blogs are on the internet by definition and they frequently link each other, they may be seen as interconnected and socially networked, through blogrolls, comments, links and backlinks. According to Weiss no technology ever led to such a revolution in "navel-gazing" as the blog in 2004 [Weiss, 2004]. Today more than 158 million identified blogs are estimated, with more than 1 million new blog posts being produced every day. This diffusion is supported by several services (such as Blogger[13] or Wordpress[14]) allowing users to create a blog in a few minutes.

3.2.5. Microblogs

Microblogging is normally defined as a form of blogging where posts, or microposts, are limited to a much shorter length than traditional blog posts. Typically, a post consists of one sentence or a Web link to a media object with a short comment. The microblogging service that made this type of social media popular is Twitter[15]. Twitter was created in

[11] Numbers offered by the WikiMedia reports at "Wikipedia Statistics" http://stats.wikimedia.org/EN/Sitemap.htm (September 30, 2013)
[12] According to Alexa (www.alexa.com, accessed January 2014)
[13] www.blogger.com (accessed January 2014)
[14] http://wordpress.org/ (accessed January 2014)
[15] http://twitter.com (accessed January 2014)

2006 and in 2012 reached 500 million registered users, posting approximately 340 million tweets per day [16] [Kwak et al., 2010].

Essentially, Twitter allows users to share messages, also known as status updates or tweets, containing maximum 140 characters. Users can keep up to date with the updates of other users by *following* them. They can also keep track of conversations by searching for topics or usernames of interest [Java et al., 2007]. Topics are normally expressed with *hashtags*, special keywords which start with the "♯" character and help in categorising the tweets. Tweets can contain mentions of user names, specified by prefixing the user name with an "@" symbol. Microblog users often *retweet* other user's tweet to share the message to their own followers [Boyd et al., 2010]. Status updates can be either public or restricted to a selected list of users. However, for this type of social media the majority of the shared messages and objects are public. As a large and publicly available source of online conversations, Twitter has become a popular source of data for researchers performing analysis of online communities. This is also demonstrated by the increasing number of research publications involving Twitter presented in the last years at international Web conferences.

3.2.6. Online Forums

An Internet Forum, or a Message Board, is a website which enables conversations about any topic. Message boards are related to earlier technologies like *Usenet*, a network of servers that enables users to post articles and reply within threads, and Bulletin Board Systems, software that allowed users to connect by terminal and exchange messages in public boards. Message boards are organised into a hierarchical structure of forums, which may themselves contain subforums. This hierarchical organisation usually follows a topic-based organisation for the forums and subforums, where each forum corresponds to a particular topic. Within a forum, a user can create a *thread*, which is a container for a single conversation, and other users can reply with follow-up posts. In order to post messages, depending on the forum's settings, users can be anonymous or have to register with the forum. Often, registered users are also organised in different groups according to their assigned privileges and rights.

[16] Twitter Blog https://blog.twitter.com/2012/twitter-turns-six (accessed January 2014)

3.2.7. Content Sharing Services

Content sharing services are a large category of social media sites enabling users to share some type or types of media such as photos, music and videos. Typically, uploaded content can be commented by the members of the community, or by a restricted set of users according to the privacy settings of the content. Users can subscribe to the updates, or feeds, of other users or can join groups related to particular topics of interest. YouTube, with more than 1 billion unique users visiting the website each month and over 6 billion hours of video watched each month[17], is the most popular video sharing service and the third most popular website globally after Google and Facebook[18]. It is one of the best examples of content sharing services, while in this case the shared media is videos. Similarly, photo sharing services such as Instagram[19] and Flickr[20] currently benefit of huge popularity and large user bases. What is common between these applications is that they not only allow comments to the media objects, but they also allow content to be annotated with titles, tags, categories and descriptions [Marlow et al., 2006]. This clearly facilitates the organisation and search of the content.

3.2.8. Social Bookmarking Services

A social bookmarking system enables its users to add, annotate, edit, and share bookmarks of Web documents [Noll and Meinel, 2007]. Users can create collections of bookmarks, share them publicly or keep them private, and access them via a Web browser anytime. By making these collections public, users allow other members of the community with similar interests to view, import and comment the links. Similarly to content sharing services, bookmarks can be organized by assigning tags or categories to each one. These sets of tags generate so called "folksonomies", which are informal ways to socially annotate and classify Web content. One of the most popular social bookmarking services, Delicious (also called *del.icio.us*), founded in 2003, pioneered folksonomies and coined the term *social bookmarking* [Mathes, 2004].

[17] YouTube Press Statistics: http://www.youtube.com/yt/press/en-GB/statistics.html (accessed January 2014)
[18] According to Alexa (www.alexa.com, accessed on January 2014)
[19] http://instagram.com/ (accessed January 2014)
[20] http://www.flickr.com/ (accessed January 2014)

3.3. Aggregation of Social Web Data

3.3.1. Vocabularies Describing Social Media

Substantial effort have been shown by the Semantic Web research community in providing a representation of the social media through the publication of standard vocabularies or ontologies. The earliest and most popular effort was the FOAF — Friend Of A Friend — project[21], followed by the SIOC — Semantically-Interlinked Online Communities — project[22]. FOAF is one of the most popular lightweight ontologies on the Semantic Web developed for representing user personal information and social relations [Brickley and Miller, 2010] (Section 3.3.1.1). While SIOC, with its lightweight ontology, was designed for the integration of online community information [Berrueta et al., 2007]. Hence, it allows the description of information contained within online community sites such as blogs, forums, wikis, etc. (Section 3.3.1.2). It has recently achieved significant adoption through its usage in a variety of commercial and open-source software applications, and is commonly used in conjunction with the FOAF vocabulary for expressing personal profile and social networking information [Graves et al., 2007]. By becoming a standard way for expressing user-generated content from such sites, SIOC and FOAF enable new kinds of usage scenarios for online community site data, and allow innovative semantic applications to be built on top of the existing Social Web. Additional modules, such as the SIOC Types and the SIOC Actions modules, have been developed to extend the capabilities of the core ontology (more details in Section 3.3.1.2).

Many other ontologies and projects have extended FOAF and SIOC for fine grained and extensive modelling of particular scenarios. For instance, *DLPO* and *Bottari*, are two examples of lightweight ontologies built on top of FOAF and SIOC. Bottari [Celino et al., 2011] is an ontology developed as part of a research project dealing in particular with microblogs and social data streams. In this specific scenario, an extension of SIOC has been implemented. In particular, the extension improves the modelling of relationships in Twitter, the connection of tweets, locations, and sentiments. DLPO (Digital.Me Live Post Ontology)[23] [Scerri et al., 2012] is built on top of popular Semantic Web ontologies, such as FOAF, SIOC, and SKOS. It models personal and social knowledge discovered from social media, it aims at interlinking posts across personal social networks. The ontology introduces some new concepts regarding: different kinds of

[21] http://www.foaf-project.org/ (accessed January 2014)
[22] http://www.sioc-project.org (accessed January 2014)
[23] http://www.semanticdesktop.org/ontologies/dlpo/ (accessed January 2014)

posts (e.g. retweets), microposts, online presence, physical presence, and online sharing practices.

Not only Semantic Web ontologies have been proposed for the representation of social media objects and activities. In particular, Activity Streams[24] is a project aiming at developing a standard protocol to syndicate activities across social media systems. The project published a vocabulary for describing social web actions and content. This vocabulary was originally designed to be serialised both in JSON format and in XML format[25] allowing activities on social objects to be expressed within the Atom Syndication Format [Nottingham and Sayre, 2005]. To note that Activity Streams has been developed and adopted by many relevant companies such as Microsoft, Google, IBM, Facebook, etc. A RDF vocabulary named *Atom Activity Streams in RDF* (AAIR), which maps the Activity Streams concepts to RDF, has been developed by the Semantic Web community[26]. The principal benefit of having a RDF vocabulary defining the core terms of this project is that they can be used in combination with other vocabularies, such as SIOC, extending their expressive limits. Activity Streams is still an ongoing project and newer revisions to the vocabulary are being developed. In this thesis we take into consideration this ample set of terms describing social media for possible extensions of SIOC and FOAF and alternative modelling solutions.

In the following subsections we provide more details about FOAF, SIOC and their extensions. These are the core vocabularies used in our work for describing and aggregating Social Web data. In this chapter we also describe how to use these vocabularies in particular practical use cases, such as in the case of wikis. We agree with the related work described previously on the fact that these ontologies are limited in describing particular scenarios in detail. Hence, we show how an extension of these vocabularies and a combination with other existing ones — supporting the reuse of Semantic Web ontologies — can be an optimal solution to this problem. In fact, for our modelling strategies, we adopted these ontologies for their simplicity and widespread adoption. FOAF and SIOC are indeed lightweight and simple ontologies, however we believe this is their strength and the key of their success, facilitating reuse and integration.

[24] http://activitystrea.ms (accessed January 2014)
[25] See both specifications respectively at http://activitystrea.ms/specs/json/1.0/ and http://activitystrea.ms/specs/atom/1.0/ (accessed January 2014)
[26] http://xmlns.notu.be/aair/ (accessed January 2014)

3.3.1.1. FOAF

FOAF is one of the most popular lightweight ontologies on the Semantic Web and using this vocabulary as a basis for representing users' personal information and social relations eases the integration of heterogeneous distributed user profiles. A FOAF profile consists of a FOAF `PersonalProfileDocument` that describes a `foaf:Person`: a physical person that has several properties describing her and holds online accounts on the Web. Some of the main FOAF properties describing users are [27]: `name`, `nick`, `phone`, `homepage`, `mbox`, etc. In Listing 3.1 we show an example of a FOAF profile. Apart from the basic contact information, we can see in the example that the person "Fabrizio Orlandi" holds an account on Twitter and that account is represented with a term (`UserAccount`) from the SIOC ontology. To note also that social relationships are expressed using the `foaf:knows` property, connecting `foaf:Person` instances together. Normally, FOAF profiles are integrated with the SIOC ontology to represent more precisely online accounts on the Social Web. For further details we suggest consulting [Brickley and Miller, 2010].

```
<foaf:PersonalProfileDocument rdf:about="">
  <foaf:maker rdf:resource="#me"/>
  <foaf:primaryTopic rdf:resource="#me"/>
</foaf:PersonalProfileDocument>
<foaf:Person rdf:ID="me">
  <foaf:name>Fabrizio Orlandi</foaf:name>
  <foaf:nick>BadmotorF</foaf:nick>
  <foaf:mbox rdf:resource="mailto:fabrizio.orlandi@deri.org"/>
  <foaf:homepage rdf:resource="http://www.deri.ie/about/team/member/
      fabrizio_orlandi"/>
  <foaf:phone rdf:resource="tel:+35391494035"/>
  <foaf:workplaceHomepage rdf:resource="http://www.deri.ie"/>
  <foaf:account>
    <sioc:UserAccount rdf:about="http://twitter.com/BadmotorF">
    </sioc:UserAccount>
  </foaf:account>
  [...]
  <foaf:knows>
    <foaf:Person>
      <foaf:name>Alexandre Passant</foaf:name>
      <rdfs:seeAlso rdf:resource="http://apassant.net/foaf.rdf"/>
    </foaf:Person>
  </foaf:knows>
</foaf:Person>
```

Listing 3.1: Example of a FOAF-based user profile in RDF/XML

[27] FOAF Specification: http://xmlns.com/foaf/spec/

Important for FOAF is also its relationship with *vCard*, a specification developed by the IETF for the description of people and organisations[28]. The vCard data format is used for "representing and exchanging a variety of information about individuals and other entities (e.g., formatted and structured name and delivery addresses, email address, multiple telephone numbers, photograph, logo, audio clips, etc.)." [Perreault, 2011]. Typically, vCard objects are encoded in its own defined text-based syntax or XML renderings. However, an equivalent representation of vCard utilizing the Semantic Web representations of RDF/OWL is provided by a W3C Working Draft[29]. Moreover, mappings between vCard and FOAF terms have been created by the Semantic Web community[30].

3.3.1.2. SIOC

SIOC (Semantically-Interlinked Online Communities)[31] aims to enable the integration of online community information [Berrueta et al., 2007]. It allows the description of information contained within online community sites (blogs, forums, wikis, etc.). By doing so, it makes it possible to connect these sites together, forming a Social Web of Data. SIOC provides a Semantic Web ontology for representing rich data from the Social Web in RDF. It has recently achieved significant adoption through its usage in a variety of commercial and open-source software applications, and is commonly used in conjunction with the FOAF vocabulary for expressing personal profile and social networking information. By becoming a standard way for expressing user-generated content from such sites, SIOC enables new kinds of usage scenarios for online community site data, and allows innovative semantic applications to be built on top of the existing Social Web. The SIOC ontology has been published as a W3C Member Submission, submitted by 16 organisations[32].

The SIOC ontology is composed of the main SIOC Core ontology, and three additional modules: Access, Types and Services. An additional module named *SIOC Actions* has been developed to represent the dynamics of online communities (more details in Section 3.3.1.2). However, it is not part of the original proposal for standardisation of

[28] IETF RFC6350 vCard Format Specification: http://tools.ietf.org/html/rfc6350 (accessed January 2014)
[29] vCard Ontology, for describing People and Organisations. W3C Working Draft 24 September 2013: http://www.w3.org/TR/2013/WD-vcard-rdf-20130924/ (accessed January 2014)
[30] http://wiki.foaf-project.org/w/FOAF_and_vCard (accessed January 2014)
[31] http://sioc-project.org (accessed January 2014)
[32] http://www.w3.org/Submission/2007/02/ (accessed January 2014)

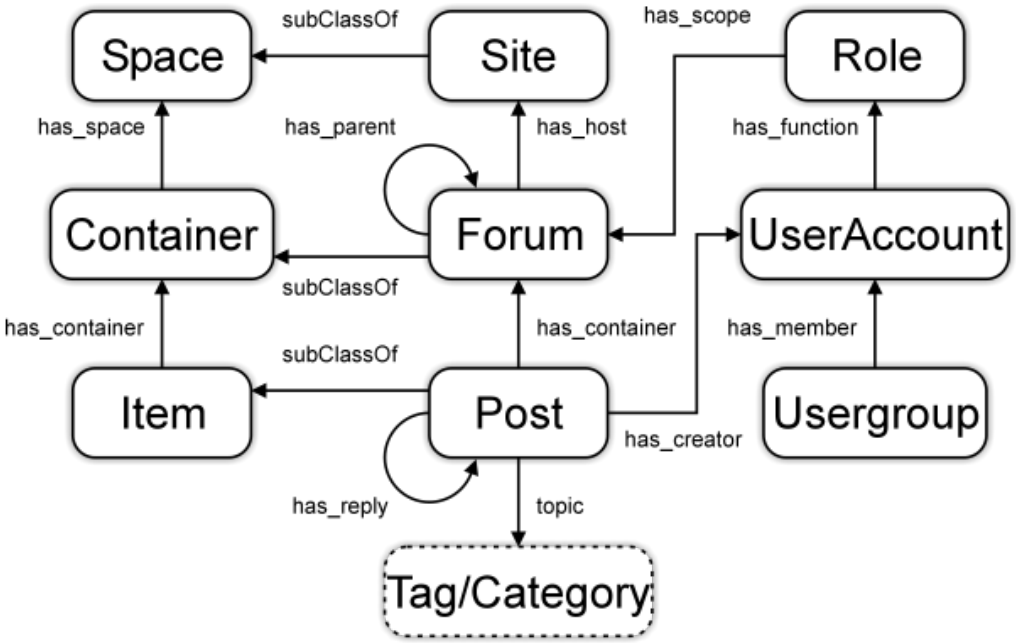

Figure 3.3.: Main SIOC Core classes and properties

SIOC. The main classes and properties of the SIOC Core ontology are illustrated in the following Figure 3.3.

An example about a very basic document describing a blog entry, taken from the SIOC Core Ontology Specification[33], is displayed in the following Listing 3.2.

```
<sioc:Post rdf:about="http://johnbreslin.com/blog/2006/09/07/creating-connections-
    between-discussion-clouds-with-sioc/">
   <dcterms:title>Creating connections between discussion clouds with SIOC</dcterms
       :title>
   <dcterms:created>2006-09-07T09:33:30Z</dcterms:created>
   <sioc:has_container rdf:resource="http://johnbreslin.com/blog/index.php?
       sioc_type=site#weblog"/>
   <sioc:has_creator>
       <sioc:UserAccount rdf:about="http://johnbreslin.com/blog/author/cloud/" rdfs
           :label="Cloud">
           <rdfs:seeAlso rdf:resource="http://johnbreslin.com/blog/index.php?
               sioc_type=user&amp;sioc_id=1"/>
       </sioc:UserAccount>
   </sioc:has_creator>
   <sioc:content>SIOC provides a unified vocabulary for content and interaction
       description: a semantic layer that can co-exist with existing discussion
       platforms.</sioc:content>
   <sioc:topic rdfs:label="Semantic Web" rdf:resource="http://johnbreslin.com/blog/
       category/semantic-web/"/>
   <sioc:topic rdfs:label="Blogs" rdf:resource="http://johnbreslin.com/blog/
       category/blogs/"/>
   <sioc:has_reply>
```

[33] http://rdfs.org/sioc/spec/ (accessed January 2014)

```
        <sioc:Post rdf:about="http://johnbreslin.com/blog/2006/09/07/creating-
            connections-between-discussion-clouds-with-sioc/#comment-123928">
            <rdfs:seeAlso rdf:resource="http://johnbreslin.com/blog/index.php?
                sioc_type=comment&amp;sioc_id=123928"/>
        </sioc:Post>
    </sioc:has_reply>
</sioc:Post>
```

Listing 3.2: Describing a blog entry with SIOC

The brief example illustrated introduces the basics of SIOC. In other words, it says:

- There is a post titled "*Creating connections between discussion clouds with SIOC*" created at *09:33:30* on *2006-09-07* written by a user "*Cloud*" on topics "*Blogs*" and "*Semantic Web*" with contents described in `sioc:content`.
- More information about its author can be found at http://johnbreslin.com/blog/index.php?sioc_type=user&sioc_id=1
- The post has a reply and detailed SIOC information about this reply can be found at http://johnbreslin.com/blog/index.php?sioc_type=comment&sioc_id=123928

This simple example uses only two classes of SIOC objects: `sioc:Post` and `sioc:UserAccount`. There are other classes in SIOC used to describe more information about users, sites, communities and other objects. Further details about the Core ontology, and a full definition of these classes and related properties, can be found in the namespace located at: http://rdfs.org/sioc/ns.
SIOC modules are used to extend the available terms and to avoid making the SIOC Core Ontology too complex and unreadable.

- SIOC Access module[34] contains classes and properties that allow to express information about access rights such as users' permissions and status of content Items.
- SIOC Types Module[35] includes some of the SIOC Core Ontology multiple subclasses for different types of Containers and Posts, such as: Wiki, WikiArticle, Weblog, BlogPost, etc.
- SIOC Services Module[36] provides a simple way to tell others about a web service (it should not be confused with web service definitions that define the details of a

[34] http://rdfs.org/sioc/access
[35] http://rdfs.org/sioc/types
[36] http://rdfs.org/sioc/services

web service). A `sioc:Service` allows us to indicate that a web service is associated with (located on) a `sioc:Site` or a part of it.

Currently, more than 50 applications are using SIOC[37], either as a common vocabulary to expose their data in RDF, alongside with FOAF for instance, as well as using existing SIOC data. By installing relevant SIOC export plugins, online community sites can generate linked data and start forming a critical mass of RDF data about user-created content [Bojārs et al., 2008]. Other tools allow users to browse SIOC data or to translate existing data, such as mailing list archives, to SIOC.

A simple and effective way to use and link to/from SIOC data is to interlink SIOC with other vocabularies such as FOAF and SKOS. By doing so it is possible to make online community data, described in SIOC, a more integrated part of the Web of Data. Common practice to facilitate linking to SIOC should be: the linking to social media sites and their user accounts on these sites by owners of FOAF profiles; then SIOC exporters can be optimized to make SIOC data easier to discover; finally Semantic Web indexing and lookup services can find and provide access to SIOC data [Bojārs et al., 2008]. A summary about the concepts of linking SIOC with other ontologies such as FOAF and SKOS, and about linking to SIOC data (especially the way to identify a user with his online accounts), is illustrated in Figure 3.4[38].

The SIOC Actions Module While SIOC represents the state of a community at a given time, SIOC Actions [Champin and Passant, 2010] can be used to represent their dynamics, *i.e.* how they evolve. Hence, SIOC provides a *document-centric* view of online communities and SIOC Actions focuses on an *action-centric* view. More precisely, the evolution of an online community is represented as a set of `Actions`, performed by a *user* with its `UserAccount`, at a specific `time`, and impacting a number of `objects`. Besides the SIOC ontology, SIOC Actions relies on the vocabulary for Linking Open Descriptions of Events (LODE)[39] described in [Shaw et al., 2009]. The core of the module is the `Action` class, which is a timestamped event involving an `agent` (typically a `foaf:Agent`) and a number of digital artefacts (class `sioca:DigitalArtifact`). Figure 3.5[40] displays a diagram with two representations of an `Action` linked to its timestamp and its actor.

[37] http://sioc-project.org/applications (accessed January 2014)
[38] Image taken from: http://sioc-project.org/node/158 (accessed January 2014)
[39] http://linkedevents.org/ontology/ (accessed January 2014)
[40] Please note that the class `sioc:User` has been renamed in `sioc:UserAccount`

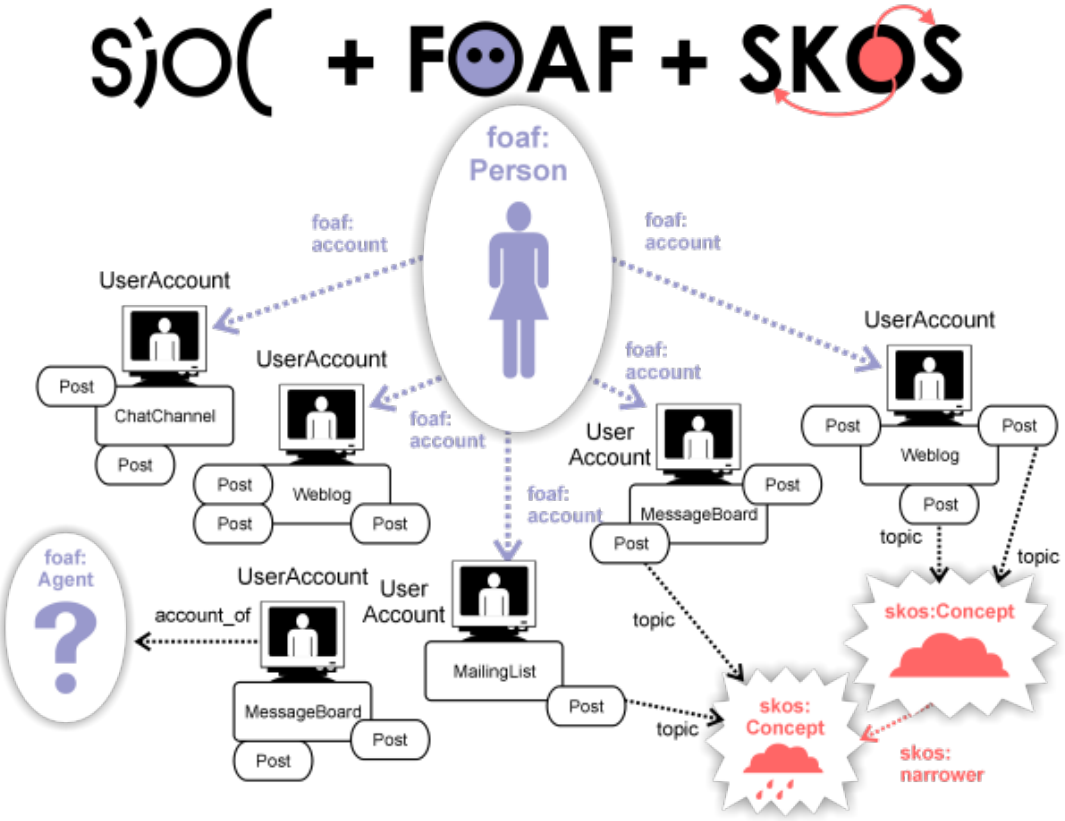

Figure 3.4.: Interlinking SIOC, FOAF and SKOS.

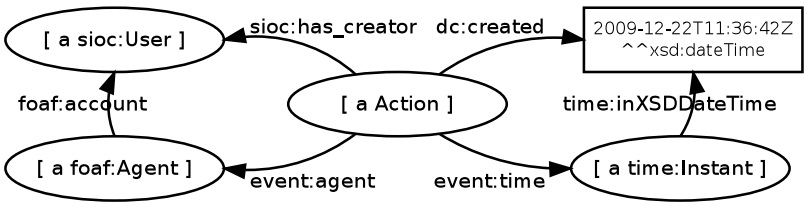

Figure 3.5.: Two representations of actor and timestamp of an action using the SIOC Actions module (Taken from [Champin and Passant, 2010])

The `Action` class is subclass of `Event` from the the Event Ontology. SIOC Actions provides an extensible hierarchy of properties for representing the effect of an action on its artefacts, such as `creates`, `modifies`, `deletes`, `uses`, etc. For a more detailed description of the implementation of SIOC Actions in a concrete example such as wikis, we invite the reader to consult Section 4.2.1.2.

3.3.2. Interlinking Social Media Systems Using Semantic Technologies

The creation of a standard and widely agreed vocabulary for describing social media systems, their users and objects allows for a structured representation of the Social Web. As thoroughly described in [Breslin et al., 2009] and the documents published by the FOAF and SIOC communities, semantic technologies provide a method for expressing information contained within online communities in a standard form. The first fundamental step in this direction is the definition of ontologies that describe the domain of online communities and what they consist of (e.g. users, activities, posts and other terms that occur in these communities). There are a lot of structures and inherent connections present in social web systems, in that people tag content, make replies or create trackbacks between posts. The structure that is created in online communities is often hidden in some database behind the scenes, and semantics can be used to expose that structure. Most of the online discussions have similar structure, whether they be on microblogs, forums or wikis. Typically, they consist of a discussion starter and some replies or comments to the original post. By using a common ontology in place, we can use this to represent and interlink the data from different communities. For example, by representing and capturing users' contributions on a wiki and aggregating this information with their posts and comments on some different blogs, it is possible to create a base for a distributed and unified view of user activities on the Social Web and use this for expertise finding, distributed conversations, cross-website recommendations, etc.

As previously mentioned (Figure 3.4), the main ontologies used for this purpose are FOAF, SIOC and SKOS (Section 3.3.1). For particular scenarios these ontologies, which represent the core structure of a modelling solution, can be refined and extended for finer-grained modelling. [Breslin et al., 2009]

Following the development of the necessary ontologies, the next step towards connecting all of these discussion primitives is encouraging people to develop and install semantic data exporters for a variety of Social Web systems. As argued in [Bojars, 2009], it is necessary to establish a complete methodology for making a common format, such as the one provided by SIOC, widely adopted. SIOC, for instance, offers a common format for expressing social media data in a rich and interlinked form. This interconnection of online communities using Semantic Web technologies can lead to many interesting possibilities both on the individual and the community level. Thanks to the data represented in a standard machine readable format, many applications and browsers taking advantage of

this information can be built on top of it. A complete "food chain" [Breslin et al., 2009] of applications needs to be deployed between data producers and consumers in order to support community engagement and the widespread adoption of the standard.

In the particular case of SIOC, in Figure 3.6 we show an example of a food chain of applications producing, collecting and consuming SIOC data (as depicted in [Breslin et al., 2009], [Bojars, 2009] and several other documents related to the SIOC project). Only a few types of applications are included in the figure illustrating where SIOC data is actually being used. Data producer applications can be natively built in applications or by directly mapping relational databases into RDF data. Alternatively, specific exporters can be developed on top of APIs or as add-ons for existing applications. The data produced can be discovered by collectors which can act as web crawlers and/or web data indexers. Finally, semantic data can be used by many applications for different purposes: to further enrich the data, for browsing, data analytics, etc.

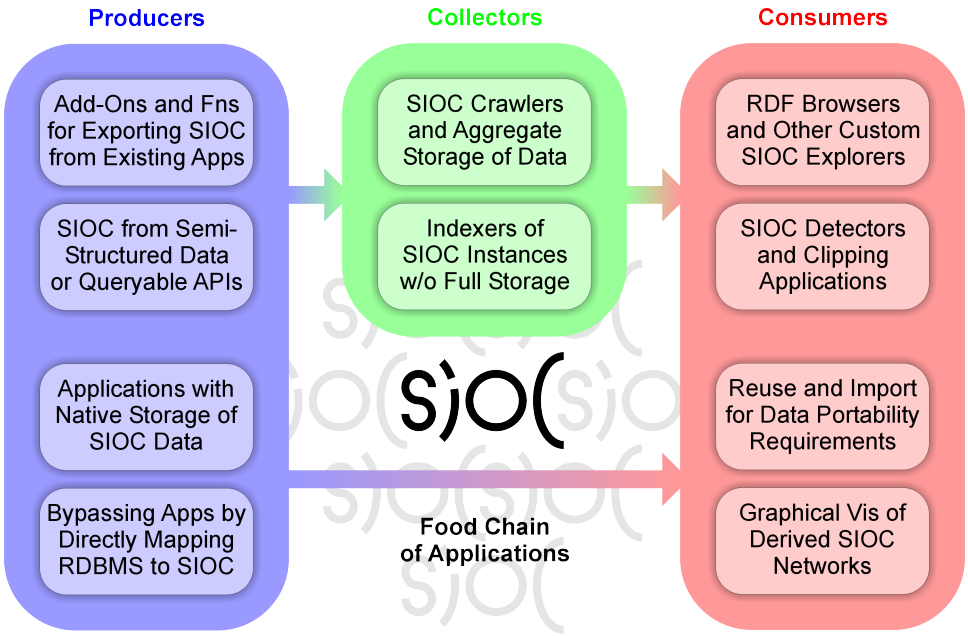

Figure 3.6.: The food chain of applications producing, collecting and consuming SIOC

In this thesis we follow the same methodology for generating Semantic Web data out of social media systems, therefore, we use it for further enrichment and data analytics. In the next section we show how we model Social Web data using standard popular ontologies and following the best Linked Data principles. Moreover, we describe possible data producers that translate data extracted from APIs, or directly from a RDBMS, into Semantic Web data following our modelling solution. In particular, in Section 3.4, we detail an application for browsing and searching on top of different wiki systems. This

represents an example of a "data consumer" which uses Social Semantic Web data in a meaningful way. It also demonstrates the advantages of this methodology compared for instance with traditional Web 2.0 approaches. In the next chapters we will describe other data consumer applications aiming at enriching and analysing Social Semantic Web data for profiling user interests.

3.4. Use Case: Enabling Search on Heterogeneous Wiki Systems

Following our previous related work [Orlandi, 2008] [Orlandi and Passant, 2009] on modelling the Social Web features of wikis as structured semantic data, in this section we describe our approach for representing and extracting Social Web data in the particular case of wikis. This particular case serves as a demonstration of the validity of the approach and can be applied to any other type of social media. The main steps described in this section for enabling an application for searching and browsing social data on different wiki systems are: (i) to model the social and structural features of a social media website using popular lightweight ontologies; (ii) to develop data producers that translate and export data according to the defined semantic model; (iii) to collect the data and build an application on top of it. These steps are in general applicable to any semantic application [Heitmann et al., 2012a]. In the following subsections we describe more practically how we built an efficient application with a simple user-interface enabling semantic searching and browsing capabilities on top of different interlinked wikis. We describe how we designed a common model for representing social and structural wiki features and how we extracted semantic data from wikis running on MediaWiki and Dokuwiki software platforms. More details are included in our publication [Orlandi and Passant, 2010].

As regards the term *semantic search*, we define it as the data searching technique that aims not only at finding relevant keywords matching an initial search query, but also at determining the intent and contextual meaning of the words (or even entities) a person is using for search. Semantic search systems leverage several different elements to provide relevant search results, such as: location, intent, variation of words, generalized and specialized queries, concept matching and natural language queries. In this particular use case we rely on retrieving knowledge from richly structured data sources represented with popular ontologies and our modelling solution, which we describe in the following Section 3.4.1. We built a faceted-browsing interface to provide users with a higher level

of expressiveness. The interface enables users to specify their intent in more detail by selecting and using entities, concepts or categories for their queries.

3.4.1. Modelling the Structure of Wikis

Typically, wikis allow editing of documents and, by definition, allow multiple users to simultaneously contribute to the content; they track history of changes so that pages can be restored to previous modified versions; they include comments or discussion areas; they link to other external sources or within the wiki; they describe categories into hierarchical structures. For each of these features, we will now describe how we modelled it, using (and extending when needed) the SIOC Core ontology[41] and its Types module[42].

Natively, the SIOC Types module already defines the `Wiki` and `WikiArticle` classes that can be used to represent the basic objects manipulated by wikis, e.g. wikis and their pages. We consequently reused these classes and added new properties to model additional features. Since this work is based on the work done before the Ph.D. studies, we will not go into detail on the modelling solution and we refer to our publications [Orlandi, 2008, Orlandi and Passant, 2009].

To summarise, the structural features modelled using SIOC, and other popular vocabularies, are the following:

- **Multi-authoring**. A fundamental feature of wikis is that multiple users are allowed to modify the same content, enabling some kind of collective intelligence process. In this regard, the semantic infrastructure should provide a model to identify users and theirs modifications, marking events with a corresponding timestamp so that provenance of information can be tracked between two versions.
- **Categories**. In many systems, wiki pages are generally related to categories, that allow readers to find sets of articles on related topics. Categories can also be organized in a tree-like structure and their semantic model should maintain the original taxonomical structure.
- **Social tagging**. While not all wiki engines support that feature, we believe this is particularly relevant, especially as it offers an open and user-driven classifica-

[41] `http://rdfs.org/sioc/ns`, prefix `sioc`
[42] `http://rdfs.org/sioc/types`, prefix `sioct`

tion scheme for wiki pages. The use of tags lead to a non-organised but dynamic organisation process, known as a "folksonomy", rather than the more widely used hierarchical structures.

- **Discussions**. Several wikis associate a discussion page to every wiki page, so that each user is able to comment and argue his point-of-view on the topic. On a discussion page, people can discuss about the article subject, or about the way that subject is presented.
- **Backlinks**. Backlinks are an important feature of wikis, as they allow to visualize instantaneously all the incoming links to a website or web page. More precisely they are wiki internal links pointing to a wiki article. It is a very common wiki feature and they may be of significant interest: they indicate who is paying attention to the linked page or topic.
- **Versioning**. Usually all editable pages on wikis have an associated page history. This history consists of the old versions of the wikitext, as well as a record of the date and time of every edit, the username or IP address of the user who wrote it, and their edit summary. All this is usually accessible through a special "history" page which shows time-ordered links to all the revisions. Commonly the latest revision of a wiki page has always the same URL (alias name), meanwhile older versions have further parameters appended to the URL (see Figure 3.7).

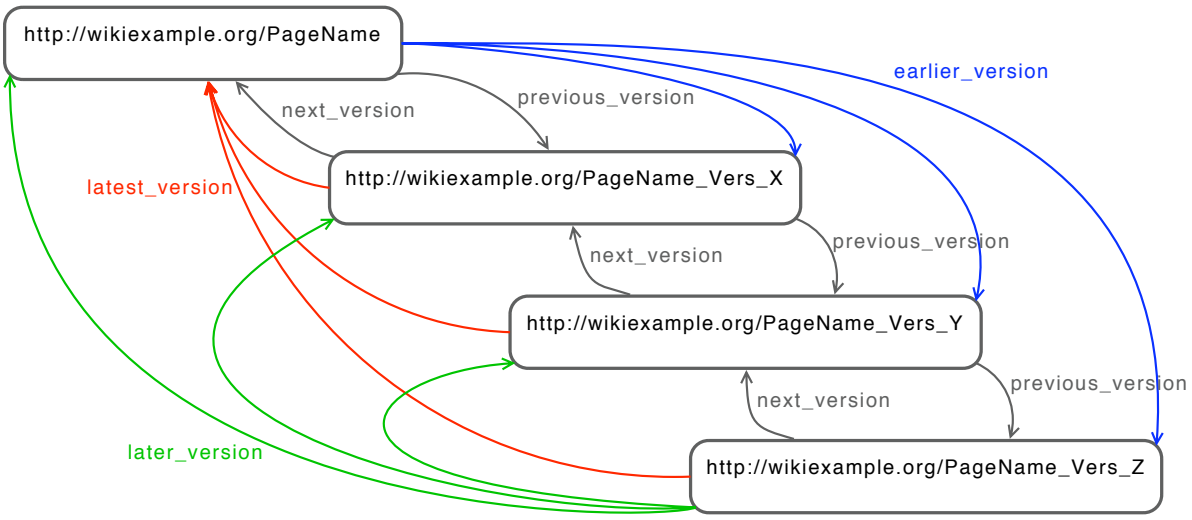

Figure 3.7.: Modelling solution for versioning of wiki articles

3.4.2. The RDF Model Generated

In this section we briefly describe the modelling solution adopted by examining the RDF data generated for a typical wiki page. We base our example on the Wikipedia page about *DERI*[43].

As mentioned before, we decided to use `dc:title`, `dcterms:created` and `dc:contributor` to model the document with the Dublin Core ontology. The choice of using `dcterms:created` to identify the date of creation of this particular revision, instead of `dcterms:modified`, has been made because the URI of a `WikiArticle` refers to a single revision. Hence, a revision could not be modified but only created. The author of any revision is a `dc:contributor`, and her username is expressed as a literal. Additionally, the user account URI, used by the author in this wiki, is modelled with a `sioc:UserAccount` class (subclass of `foaf:OnlineAccount`).

```
<sioct:WikiArticle rdf:about="http://en.wikipedia.org/wiki/
    Digital_Enterprise_Research_Institute">
 <dc:title>Digital_Enterprise_Research_Institute</dc:title>
 <foaf:primaryTopic rdf:resource="http://dbpedia.org/resource/
     Digital_Enterprise_Research_Institute"/>
 [...]
 <dc:contributor>StefanDecker</dc:contributor>
 <dcterms:created>2008-12-11T12:59:19Z</dcterms:created>
 <sioc:topic>
   <sioct:Category rdf:about="http://en.wikipedia.org/wiki/Category:
       Scientific_organizations">
   [...]

 <sioc:links_to>
   <sioct:WikiArticle rdf:about="http://en.wikipedia.org/wiki/Semantic_Web">
   [...]

 <sioc:has_discussion>
   <sioct:WikiArticle rdf:about="http://en.wikipedia.org/wiki/Talk:
       Digital_Enterprise_Research_Institute">
   [...]

 <sioc:has_container>
   <sioct:Wiki rdf:about="http://en.wikipedia.org"/>
 [...]
</sioct:WikiArticle>
```

Listing 3.3: Example of a RDF description of the Wikipedia page about DERI

The above Listing 3.3 has been reduced for displaying purposes, every resource has a `rdfs:seeAlso` link associated (as illustrated later in Listing 3.5 for the description of a

[43] http://en.wikipedia.org/wiki/Digital_Enterprise_Research_Institute

`sioc:UserAccount`) and, of course, a closing tag. As we can see, categories are mapped into the `sioct:Category` class through the `sioc:topic` property. Internal and external links are described with `sioc:links_to` links. Discussion pages are marked in Wikipedia with the "Talk:" prefix[44] and defined in range of the `sioc:has_discussion` property. The `sioct:Wiki` container, which identifies the wiki site hosting the `sioct:WikiArticle`, is expressed with the `sioc:has_container` property.
Pages versioning, is modelled in RDF as illustrated in the example of Listing 3.4.

```
<sioc:previous_version>
  <sioct:WikiArticle rdf:about="http://en.wikipedia.org/w/index.php%3Ftitle%3
      DDigital_Enterprise_Research_Institute%26oldid%3D246494912">
  [...]

<sioc:latest_version>
  <sioct:WikiArticle rdf:about="http://en.wikipedia.org/wiki/
      Digital_Enterprise_Research_Institute">
  [...]
```

Listing 3.4: Versioning for Wikipedia articles

In Listing 3.4 we note that the URI of a previous version is typically marked by MediaWiki with the "`oldid`" parameter appended. Furthermore, Listing 3.4 represents an export example of the latest version of the "Digital_Enterprise_Research_Institute" article, hence, a newer `sioc:next_version` does not exist yet. An illustration to better summarise our versioning model for wiki pages is in Figure 3.7.

This section provided an overview of the RDF modelling solution adopted for wiki pages, the next sections describe our approach and implementations for generating RDF data out of wikis following our presented model.

3.4.3. Exporting SIOC Data From Heterogeneous Wikis

Following the definition of a common interchange model for wikis, in order to evaluate our proposal, we decided to generate and collect a substantial amount of structured data in RDF/XML format, generated from different wiki platforms. First, a webservice that exports every wiki page from the MediaWiki software platform in RDF has been developed (Section 3.4.3.1). This exporter is called "SIOC-MediaWiki exporter" and it is publicly available on the Web. Our attention was focused on the MediaWiki platform simply because it is one of the most popular wiki platforms on the Web, hosting all the

[44] http://en.wikipedia.org/wiki/Wikipedia:Talk_page

Wikimedia Fundation wikis (i.e. Wikipedia, Wiktionary, etc.) and propulsing more that 69 millions of wiki articles from different wiki sites[45].

The second wiki platform we chose is DokuWiki[46], which is another popular wiki platform together with TWiki and MoinMoin, aimed at small companies' documentation needs and particularly suited for fast and easy setups and configurations since it does not need a database. We focused on this wiki software also because a plug-in for DokuWiki that exports RDF data, and especially SIOC ontology based data, has been already developed by Michael Haschke[47], a contributor of the SIOC project's community. However, it was not fully-compliant with our proposed model and some features we needed were missing, therefore we adapted and improved this plug-in implementation in order to meet the requirements we established (Section 3.4.3.2).

Exporting and collecting data extracted from two different and relevant wiki platforms allows us to evaluate and demonstrate the validity of our approach. It gives us the possibility to run and experiment cross-wikis and cross-platforms queries and to show some of the potentialities of Semantic Web technologies, as we will show in Section 3.4.4.

As regards the techniques to continuously update the Social Web information collected, we can use live updates or feeds (e.g. RSS feeds) to continuously update our database and keep it "fresh" after the first complete crawl of the site.

3.4.3.1. The SIOC-MediaWiki Exporter

The SIOC-MediaWiki webservice is written in PHP and is publicly available at `http://ws.sioc-project.org/mediawiki/`. It exports any MediaWiki wiki article in RDF using the structure explained previously. This work was also part of our previous work, hence, for more details we refer again to our publications [Orlandi, 2008] [Orlandi and Passant, 2009].

To briefly explain the characteristics of the exporter, it is relatively lightweight and built using only two PHP classes: the SIOC-MediaWiki exporter itself and the already existing SIOC API[48]. The latter has been improved in order to take the new characteristics of the model into account. The exporter class is the part responsible for querying the MediaWiki API and parsing the results, and the SIOC API is responsible for exporting

[45] `http://s23.org/wikistats/largest_html.php` as of November 2013
[46] `http://www.dokuwiki.org/` (accessed January 2014)
[47] `http://eye48.com/dokuwiki/doku.php?id=en:dokuwiki:sioc-plugin` (accessed January 2014)
[48] `http://wiki.sioc-project.org/index.php/PHPExportAPI` (accessed January 2014)

the content in RDF. The script automatically discovers the MediaWiki API location of the requested wiki, then it connects to the API with HTTP GET requests as queries. After parsing the results of the queries it calls the SIOC API to export in RDF/XML serialization the fetched structural information.

Since the initial release of the exporter, we focused on improving the performances of the application, especially in terms of response time. This is a very important requirement: considering the process of crawling a wiki using the exporter, even a small reduction of the time needed to export a single wiki page would lead to a consistent amount of time saved when collecting data for all the pages in an entire wiki. Unfortunately the exporting time with the SIOC-MediaWiki webservice is strongly dependent on (i) the time of response of the API of the original MediaWiki system that is exported and (ii) on the number of queries needed to get all the data. The second aspect, on which we concentrated our attention, was the way users and anonymous users are modelled. In particular the anonymous users need to be modelled and they can only be linked to a blank node. Finally the possibility to finely select the relevant wiki structural features to export has been added. Then, users can decide to export only some basic information on a wiki article and ignore other information, for instance exporting revisions but no categories, instead of being always forced to export them all. This enables better usage of the applications, as third-party developers can concentrate on extracting only the required subset of the original systems.

3.4.3.2. The DokuSIOC Plugin for DokuWiki

The main functionalities offered by DokuWiki are extensible by implementing plugins, called Action Plugins, which are designed to work with DokuWiki events to allow for customization of any part of DokuWiki that signals its activity using events. The DokuWiki documentation[49] gives detailed information on their structure and how to develop new plugins. In this section, we focus more on DokuSIOC, a plugin developed by Michael Haschke[50], and how we extended it to fit with our SIOC extensions. Action plugins are loaded before any significant DokuWiki processing takes place. At load time they register their event handlers so that when a specific event is signalled all event handlers registered for that event are called. Hence, plugins have the opportunity to alter either the event data or the event's subsequent processing.

[49] http://www.dokuwiki.org/devel:action_plugins (accessed January 2014)
[50] http://eye48.com/go/dokusioc (accessed January 2014)

The DokuSIOC plugin takes information from the metadata stored in the wiki system about pages, users, links, and revisions and provides it as raw RDF/XML serialized data (instead of the usual HTML page) if asked for it. Futhermore, DokuSIOC provides several different ways to offer its service to clients. A simple way is to add the GET parameter `do=export_siocxml` to the URL of a wiki page, or to follow the meta link added to the header of the DokuWiki HTML view. Another option is based on the content-negotiation capability: if the client requests the usual URL of the page with an HTTP header asking for *application/rdf+xml*, the plugin will forward to the location of the RDF export view. These options are particularly useful as regards the crawling process of a DokuWiki wiki using a common RDF crawler which can automatically discover the linked RDF data.

The semantic model used by the DokuSIOC plugin after our modifications, reflects exactly the model we detailed in the previous sections, as well as the one used by the SIOC-MediaWiki exporter. One of the problems encountered when developing this DokuWiki plugin relates to the internal handling of user identifiers and profiles in it and consequently how to model URI for users. In this case the DokuSIOC plugin was already offering a way to configure a DokuWiki namespace, where user identifiers can be used as sub pages. The following URI structure `http://[dokuwikiurl]/doku.php?id=user:username` provides the identifier for the user account on the wiki. Moreover, a usual DokuWiki URL can stand for different resources, any URL may describe either a user as a `sioc:UserAccount` or a wiki page as a `sioct:WikiArticle` or a container (in this case a specific `sioct:Wiki` wiki container is more appropriate). In this regard, the SIOC plugin adds a type parameter to distinguish exactly between the resources types. Different URI structures are then used depending on the context, e.g. `http://[dokuwikiurl]/doku.php?id=user:username&type=user` for a user and `http://[dokuwiki]/doku.php?id=pageid&type=post` for an article.

To generate RDF data out of the metadata extracted from the wiki system and easily create SIOC documents, DokuSIOC uses the SIOC PHP API similarly to the MediaWiki exporter previously described. An important change we have made in the plugin implementation has been to use the SIOC PHP API as much as possible to create the SIOC objects such as the `sioct:WikiArticle`, the `sioct:Wiki` container, the `sioc:UserAccount` etc. In the previous implementation of the plugin there was a PHP class that was acting as a mediator between the main `Action` class of the DokuWiki plugin and the `sioc_inc.php` script of the SIOC API. This "intermediate" class was used to change and customize the behavior of the SIOC API in order to create per-

sonalized objects using the methods provided by the SIOC API. In our perspective the SIOC API gives us all the instruments and objects we need in order to have the same wiki modeling between the different wiki platforms and to keep the interoperability between them. Hence, we decided to keep the same structure we used for the SIOC-MediaWiki exporter and relay on the SIOC API implementation[51]. In particular our changes have been focused on the properties used to define the contributors of the articles (we use the `sioc:has_creator` to point to the `sioc:UserAccount` and the `dc:contributor` for the username as literal), and the date of creation of each article revision (with `dcterms:created`). Furthermore, we changed the way the backlinks were modeled deciding to keep the same `sioc:links_to` property used for forward links. This particular choice would ease the querying part of our work because we have always the same property expressing links between the articles, no matter if they are back/forward or internal/external links.

Another relevant contribution we made to the DokuSIOC plugin was to add the "external links" feature which was not implemented. Indeed, DokuWiki does not provides native metadata about the external links linked from each page. Our exporter consequently parses all links from HTML articles and extract the external ones. Once the extraction has been made by the "Action" main class of the plugin, we export them using the same criteria as the internal links.

3.4.3.3. Following Linked Data Principles

The main goal of our work with the implementation of two different exporters sharing the same data model was not only to create RDF data from any MediaWiki or DokuWiki page, but also to easily allow interlinking between various wiki platforms, as well as between wiki data and other RDF data, whatever it is social data modeled with FOAF or SIOC or any other kind of RDF data. To do so, we followed the Linked Data principles defined by [Berners-Lee, 2006a] and the related best practices [Ayers and Völkel, 2008] [Bizer et al., 2007]: (i) use URIs as names for things; (ii) use HTTP URIs so that people can look up those names; (iii) when someone looks up a URI, provide useful information, using the standards (RDF, SPARQL); (iv) include links to other URIs so that they can discover more things.

[51] Another advantage of relying on the API is that any changes on the SIOC Ontology are immediately replicated in the API. Then, the DokuWiki plugin (as well as the MediaWiki one) are constantly up-to-date with the ontology changes, with only a few efforts (simply loading the new API version in the exporters).

Particularly, to offer a better browsing experience and ease the process of crawling SIOC exports of MediaWiki instances, our webservice automatically produces `rdfs:seeAlso` links between wiki pages. Actually, more than a simple link to the wiki page, the exporter provides a link to the related RDF document, as we can see in Listing 3.5 related to the export of a particular `sioc:UserAccount`. In the example, we distinguish the concept itself (i.e. `User:StefanDecker`) and the related RDF page.

```
<sioc:UserAccount rdf:about="http://en.wikipedia.org/wiki/User:StefanDecker">
 <rdfs:seeAlso rdf:resource="http://ws.sioc-project.org/mediawiki/mediawiki.php?
     wiki=http://en.wikipedia.org/wiki/User:StefanDecker"/>
</sioc:UserAccount>
```

Listing 3.5: User Modeling in the MediaWiki exporter

These `seeAlso` links are very useful not only to provide link to other related RDF documents, that can be used for instance when browsing data with Tabulator, but also in a crawling perspective. A RDF crawler could easily follow all the `seeAlso` links found on every document and continue to crawl. In this regard, for example, we crawled and exported entire wiki sites just following these links. A different approach, but with the same scope, has been adopted with the DokuSIOC plugin. As described in the previous section using content-negotiation it is possible to switch between the standard HTML view of the wiki article and its RDF representation, moreover, a meta link added to the header of the DokuWiki HTML view points to the semantic representation of each article easing the RDF data discovery process.

In a Linking Open Data perspective a relevant opportunity is the association between the wiki user's `OnlineAccount` and the `foaf:Person` holder of the account. And this is possible with the `foaf:holdsAccount` property. Using this feature it becomes possible to interlink precisely all the user accounts on different wikis belonging to the same person and then, for example, to know what are the contributions made by the same persons on different wikis, what are their interest areas, etc. At the moment it is possible but since most of the wiki users do not provide their FOAF profile, we still have to use the username as a literal, with all the ambiguities and inaccuracies that this method brings.

Another interesting feature is the linkage to the corresponding DBpedia resource (DBpedia being the RDF export of Wikipedia, Sec.2.1.2.4), if the article belongs to the English Wikipedia. Since DBpedia semantically models the content of a Wikipedia page, this connection is very useful to link semantic data about the content and the structure of a wiki article. DBpedia resource URIs are used in range of the `foaf:primaryTopic` property, as this property relates a document to the main thing that the document

is about. Obviously this linkage between DBpedia and Wikipedia is immediately possible only with the MediaWiki exporter, since Wikipedia is based on the MediaWiki software. However, a future improvement could be on topic extraction from pages belonging to other wikis, so that it would be possible to link every wiki page to the related Wikipedia/DBpedia categories or even to corresponding similar articles, enabling better interlinking capabilities across wikis.

3.4.4. Application for Cross-wikis Semantic Search

In this section, we will detail how we designed a Semantic Web-based application using semantic data generated from the previously detailed systems. In particular our main objective is to show that the wiki model we propose allows for interoperability between wiki platforms, and that Semantic Web technologies can (i) really improve our usual wiki experience based on typical Web 2.0 applications and (ii) permit to discover new knowledge in a faster and more accessible way.

As a first step, we exported and crawled different MediaWiki and DokuWiki instances. Five different wikis have been crawled, four from the MediaWiki platform and one from the DokuWiki one. Each MediaWiki site has been crawled using a single entry point thanks to the use of the `rdfs:seeAlso` links. We used only one entry point (or "seed") for the crawling of each wiki as our aim was not to obtain complete images of the wikis, instead our aim was to get a representative sample for our experiment. The DokuWiki wiki has been installed locally and a subset of the data from the official PHP wiki has been imported in it[52] (since our DokuWiki plug-in is not implemented in that wiki). It is important to note that each wiki we crawled belongs approximately to the same area of interest in order to have a high probability of shared topics and users. The MediaWiki sites collected are: *Semanticweb.org*[53], *Protégé Wiki*[54], *RDFa Wiki*[55] and the *ONTOLORE Karlsruhe* wiki[56], all focusing on Semantic Web technologies, with shared contributors as we will see next.

In total, we collected about 1GB of RDF data and loaded it in the OpenRDF Sesame [Broekstra et al., 2002] triple-store (Section 2.1.2.3). As we needed an higher degree of inference (because we use OWL transitive properties in our model) we also installed

[52] The official PHP.net wiki: http://wiki.php.net/ (accessed January 2014)
[53] http://www.semanticweb.org (accessed January 2014)
[54] http://protegewiki.stanford.edu (accessed January 2014)
[55] http://rdfa.info/wiki/RDFa_Wiki (accessed January 2014)
[56] http://logic.aifb.uni-karlsruhe.de/wiki/ONTOLORE (accessed January 2014)

and configured the reasoning engine OWLIM[57] on the top of it. The crawling process of all the wikis took about one entire day (24 hours), and every operation has been made on only one single-core machine. In total we collected around 45,500 triples, 3,400 wiki articles and 700 users. Once all the data has been collected it has been inserted in a Sesame+OWLIM triple-store. This process, because of the OWL inference (new triples are entailed at loading time in Sesame+OWLIM), took around two hours to be completed on the same machine, but then every query ran with the SPARQL endpoint did not take more than 3 seconds to be executed, in spite of the complexity of some of them, as we will see. As regards the scalability of the system our implementation is completely independent by the underlying triple-store. Several RDF stores have been demonstrated as capable to address the scalability requirement with a large amount of data. A comprehensive study, and a benchmark experiment, comparing the performance of popular RDF stores has been conducted in [Bizer and Schultz, 2011].

After this configuration step, the system was ready to be tested with SPARQL queries. In the following section some of the advanced queries we ran are detailed. Then, in Section 3.4.4.2 we will describe the structure of the application for semantic search and faceted browsing we built on top of the triple-store and its SPARQL endpoint.

3.4.4.1. Advanced Querying and Cross-Wiki Integration

Since our data has been loaded in an RDF store, all the queries were done using SPARQL (Section 2.1.2.3). As we can see, it offers the advantage of having a single and standard language to query wiki data, while developers that need to query original systems have to learn a new API for each new system we want to query. Then, we solved one issue that we mentioned originally in our motivation, i.e. the problem of having different ways to query different wikis.

```
SELECT DISTINCT ?wikiArt ?Contrib_a ?Contrib_b
WHERE {
 ?x sioc:latest_version ?wikiArt.
 ?wikiArt sioc:earlier_version ?VersA .
 ?VersA sioc:earlier_version ?VersB ;
  dc:contributor ?Contrib_a .
 ?VersB sioc:earlier_version ?VersC ;
  dc:contributor ?Contrib_b .
 ?VersC dc:contributor ?Contrib_a .
 FILTER (?Contrib_a != ?Contrib_b) .
}
```

[57] http://www.ontotext.com/owlim/ (accessed January 2014)

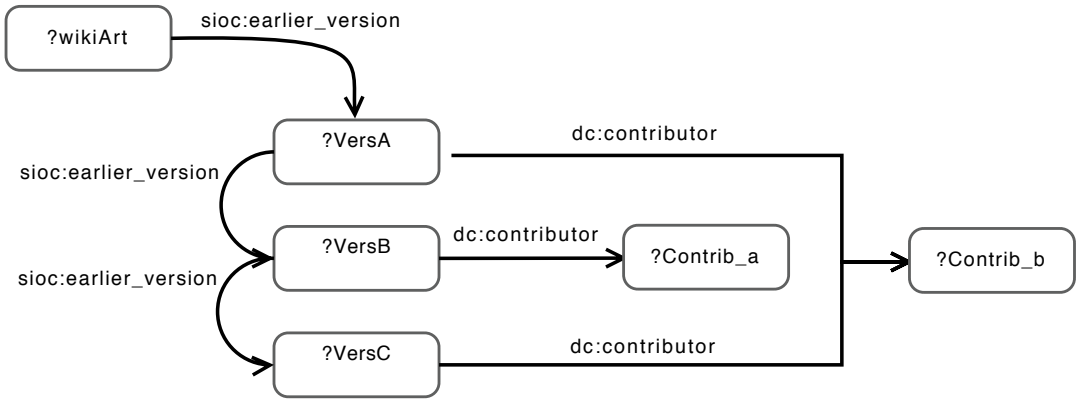

Figure 3.8.: Identifying collaborating users with SPARQL

Listing 3.6: Identifying collaborating users

A first example of advanced querying for a particular wiki is the ability to answer to the following question: "what are the collaborating users that worked alternatively on the same wiki article?". In Listing 3.6 we provide the SPARQL implementation of this query, while in Figure 3.8 we display a diagram that summarizes it. As we can see, this query takes advantage of the transitivity of the property `sioc:earlier_version`, since we identify users that worked on earlier versions, and not only immediately on the previous one.

The query provides the article URI and the two usernames in case the first user (`?Contrib_a`) re-edited the article after a modification made by the second user (`?Contrib_b`). It enables people to look for users sharing the same interests and knowledge areas. It can be also very important especially in a Social Semantic Web context.

Another interesting feature of our approach is the ability to do cross-wikis querying, since wikis are now based on the same model. The following query, in Listing 3.7, identifies users involved in different wikis, looking for the same usernames.

```
SELECT DISTINCT ?creator1 ?page1 ?page2 ?wiki1 ?wiki2
WHERE {
 ?page1 sioc:has_container ?wiki1 ;
  dc:contributor ?creator1 .
 ?page2 sioc:has_container ?wiki2 ;
  dc:contributor ?creator2 .
 FILTER (str(?creator1)==str(?creator2)) .
 FILTER (str(?wiki1)!=str(?wiki2)) .
}
```

Listing 3.7: Identifying pages created by a single user in different wikis

Yet, as this query relies on a FILTER clause, it will identify common users only if they use the same account name on two different wikis. Moreover, we can imagine that some common account names will be used by different people on different wikis, e.g. JohnSmith. To that extend, we can benefit from the strong ties that exist between FOAF and SIOC and the fact that we are modeling a wiki user using the `sioc:UserAccount` class. One person can indeed define in his FOAF profile the various wiki accounts he owns, using simple `foaf:holdsAccount` properties. Then, the previous query can be adapted to deal not only with text strings to identify the user, but with their related accounts from the FOAF URI, so that a single query can be used to retrieve all the contributions of a user whatever the wiki used was. Moreover, since the wiki model is based on SIOC, the same query can be used to retrieve wiki pages, blog posts, etc. as follows.

```
SELECT DISTINCT ?content
WHERE {
 <http://example.org/js#me> foaf:holdsAccount ?account .
 ?account rdf:type sioc:UserAccount .
 ?content sioc:has_creator ?account .
}
```

Listing 3.8: Cross-site query using FOAF and SIOC

3.4.4.2. Enabling Semantic Search

As described at the beginning of Section 3.4, we decided to build a faceted-browsing application to provide users with a higher level of expressiveness while searching on top of the described social semantic data. The interface enables users to specify their intent in more detail by selecting and using entities, concepts or categories for their queries. The application we built — to show the potential of semantic technologies applied to wikis — has the typical architecture of many Semantic Web applications. Its structure can be divided in three layers concerned with storage, querying or data acquisition, and visualization. In the previous sections we already described the storage part of the system: it is based on a Sesame+OWLIM triple-store with the data we crawled from different wikis, and it exposes a SPARQL endpoint where is possible to have an interface with the querying and acquisition module.

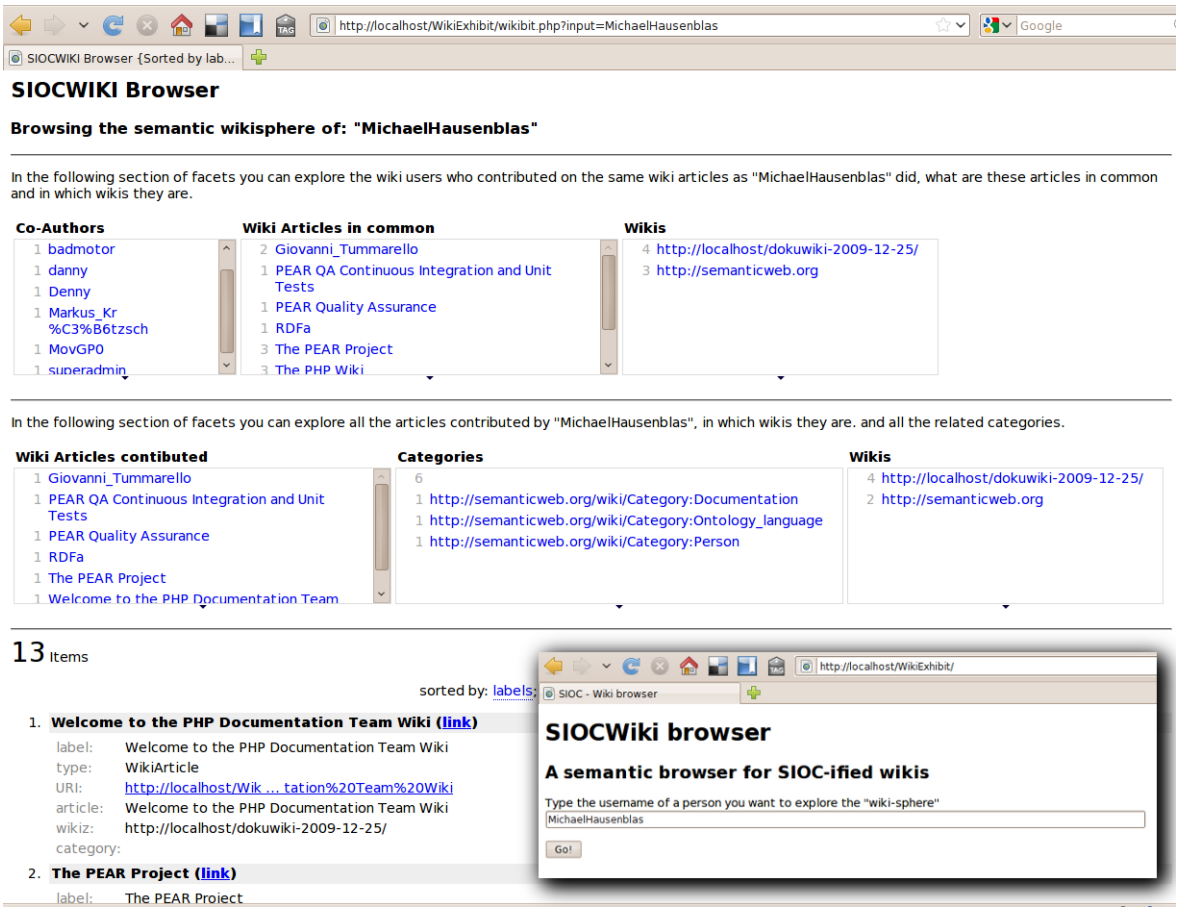

Figure 3.9.: SIOCWiki Browser: a screenshot showing the results found for the username "MichaelHausenblas".

As regards the data acquisition module we wrote a PHP script that queries our triple-store, collects and parses the results and translates the data in the correct format for the visualization layer. The PHP script is the core of the application, and in this specific application it basically needs to run two different SPARQL queries to obtain the necessary data, but it can be personalized very easily with regard to the particular desired use case.

The visualization layer has been built with the SIMILE Exhibit framework[58]. This framework allows developers to create (X)HTML pages with dynamic exhibits of data collections which can be searched and browsed using faceted browsing capabilities. Exhibit is a set of Javascript files that run in a user's browser. All it needs is a graphical configuration and personalization made directly on the HTML code of the page to display and to receive data built with a correct structure and a supported format. The

[58] http://www.simile-widgets.org/exhibit/ (accessed January 2014)

most used format with Exhibit is JSON and in our specific case this is what we adopted. In this regard our PHP script converts the XML data returned by the SPARQL queries into the JSON format.

Once the username of a wiki user has been introduced in the first page, the application provides two different informative sections. The first one is about all the wiki users who contributed on the same wiki articles as the requested user did. In other words it looks for her co-authors distributed on several different wikis. The second one provides details about all the articles contributed by the user in every wiki and the related topics of interest.

In Fig. 3.9 we display a screenshot of the developed web application. As we can see from the image, in the first horizontal section from the top there are three lists (or facets) showing the co-authors with the related wiki articles in common and the list of wikis on which the articles are located. Every element of the facets is selectable and once selected it filters all the other results on the other facets. The first section of results is obtained by the first query formulated by the PHP script. The SPARQL query used in this case is displayed in Listing 3.9 and it selects the wiki site, the wiki article and the related co-author of the user "MichaelHausenblas".

```
SELECT DISTINCT ?wiki ?title ?coauthor
WHERE {
  ?pag1 dc:contributor ?me. FILTER regex(?me, "MichaelHausenblas", "i").
  ?pag1 dc:title ?title ;
    sioc:has_container ?wiki .
  ?pag2 dc:title ?title2 . FILTER regex(str(?title), str(?title2)).
  ?pag2 dc:contributor ?coauthor . FILTER ((?coauthor) != (?me)).
}order by ?wiki
```

Listing 3.9: First query of the application

The second section of results, obtained by the second SPARQL query, displays all the articles contributed by the searched user on different wiki sites. It also adds a list of the categories (in the range of the `sioc:topic` property) related to each wiki article extracted. In other words this particular view highlights the activities, the interests and the expertise areas of the searched user. The query formulated by the script for this section is displayed in the following Listing 3.10.

```
SELECT DISTINCT ?wiki ?title ?category
WHERE {
  ?pag1 dc:contributor ?me. FILTER regex(?me, "MichaelHausenblas", "i").
  ?pag1 dc:title ?title ;
    sioc:has_container ?wiki ;
    sioc:topic ?category.
```

```
}ORDER BY ?wiki
```

Listing 3.10: Second query of the application

The last feature the SIOCWiki browser shows is a dynamic list displaying all the results extracted by the previous two sections. The results here are more detailed and they can be easily grouped and sorted. They are also filtered by the events triggered by the facets above.

3.4.4.3. Advantages of the Semantic Web Approach Compared to the Original Web 2.0 One

The Semantic Web approach showed with this application can be compared to the currently widely adopted Web 2.0 approach. Following the Web 2.0 way, in order to obtain similar results and functionalities, we would have to use each software platform separately. For example, to obtain the list of all the co-authors of one particular user we would have to: first, go to the page of the user in each wiki platform; second, use some special service provided by the wiki software to obtain her or his contributions; third, for each contribution, retrieve the history and identify all users. In addition, that workflow assumes that the wiki service provides the list of the contribution for every user, which is true for the MediaWiki platform but not for the DokuWiki one. Then, we not only simplified the process (MediaWiki) but also added some features that could not have been provided with the original tool.

Another option would be to develop some platform-specific applications which use the specific wiki software API. Once again, the interoperability is lost together with the cross-wiki global view of the data. Hence, we might state that the Web 2.0 approach can still be an option for use cases where the cross-platform interoperability is not needed and the number of the queries is limited, since these services are already available on the Web and do not require to build an infrastructure as ours. On the other hand, the Semantic Web approach needs initially more time to set-up the system (notably because of crawling and storing data) but then allows for advanced and fast querying processes and hidden knowledge discovery. It is also particularly suited for use cases such as the one we exposed with this work, namely to build an application that can be accessible to everyone and easily customizable and integrating data from different sources, based on different platforms.

3.5. Conclusions

In this chapter we provided the basis for our complete methodology for profiling user interests. We first provided a characterisation of social media and described the different Social Web activities users currently perform for interaction and content/interests sharing. Then, we provided an overview of the main vocabularies and standards for representing Social Web content, users and their actions. The semantic representation of social media is the necessary ingredient for creating a structured and interoperable meta-layer of Social Web data that can be used to aggregate user information and mine user interests. In this regard, we detailed our modelling solution for social media, which adopts several popular Semantic Web ontologies, mainly FOAF and SIOC. We described a practical experiment that applies our semantic model to a system integrating Social Web data from different heterogeneous sources. The system allows for browsing and searching capabilities on top of data collected from different wiki sites and demonstrates the validity of our modelling solution, as applied to real Social Web data. Therefore, we provided the first necessary steps for our methodology for profiling user interests: from the collection of Social Web data to its semantic representation and aggregation. The next chapter shows the integration of Social Web data with its provenance information. In Chapter 5 we detail aggregation and mining of user interests on top of this structured Social Web data layer.

Chapter 4

Provenance as Core of User Profiling Heuristics

4.1. Introduction

In Section 2.2 we provided an introduction to provenance of data, its definition and characteristics, especially in the case of the Web. In the previous chapter we argue that aggregation, and semantic representation, of Social Web data from different sources is beneficial for end-users and for the development of novel Web applications. It eases the process of aggregating and managing heterogeneous data sources and helps in providing more complete information about social media users' interests and activities. In this chapter we describe how provenance of data plays a crucial role in social media and the Web of Data. In particular we show how provenance of data can be recorded and represented on the Social Web, and consequently used on the Linked Data cloud to track the origins of particular statements and data records. At the same time, provenance on and for the Web of Data can be used in many different use cases supporting Social Web users, for user profiling, trust, data quality, etc. Therefore, we represent provenance of data as a fundamental connection between the Social Web and the Web of Data, as depicted in Figure 4.1. Thanks to provenance, a feedback loop can then be established between the two Web areas.

Provenance on the Social Web allows the Web of Data for quality control and more accurate tracking of the origins of datasets and statements. Similarly, applications built on the Web of Data would considerably benefit of more detailed and complete infor-

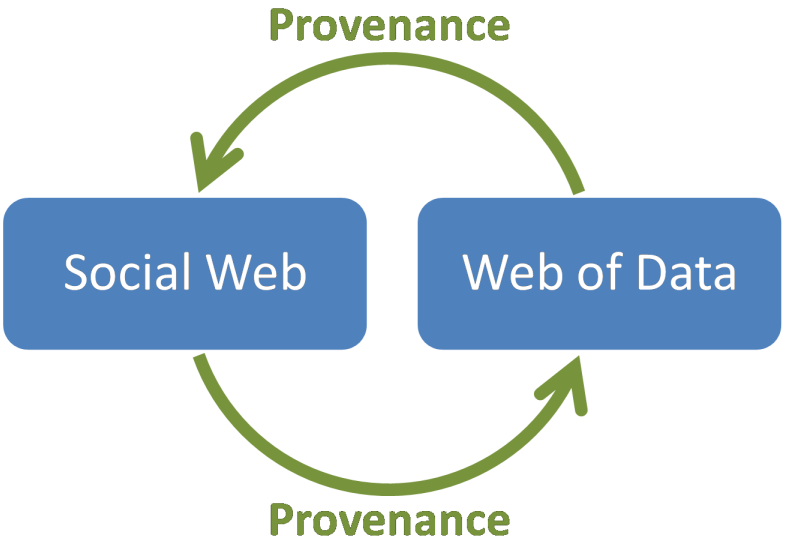

Figure 4.1.: The feedback loop between Social Web and Web of Data through provenance

mation describing the history of the datasets and their contents. Applications using information from various Web of Data datasets could use provenance information on both the dataset and the statement level to elaborate quality and trust measures. Moreover, provenance modelled using Semantic Web technologies provides clear advantages to typical Social Web scenarios such as: user profiling, recommendations of content and people, expert finding, citizen sensing and incident reporting through social media, risk management, computation of trust and reputation, etc. The main advantage in these cases is the possibility to interchange provenance, as different provenance-aware systems would natively adopt their own model for representing their provenance, but a core provenance data model would be readily adopted as a provenance interchange model across such systems.

In this thesis we demonstrate the capabilities of semantic representation of provenance in profiling user interests. In particular, in this chapter we focus on a use case involving Wikipedia and DBpedia. This use case, as described above, is a perfect example of interconnection between Social Web and Web of Data, and serves as an example for describing a more general modelling solution for provenance on the Social Semantic Web. We propose a methodology to semantically represent information about provenance of data in DBpedia and an extraction framework capable of computing provenance for DBpedia statements using Wikipedia edits (Section 4.2.1). Then, by indicating by whom and when a triple was created (or contributed by), we let any application evaluate DBpedia statements based on particular criteria (Section 4.3.1).

4.1.1. Scenario and Related Work

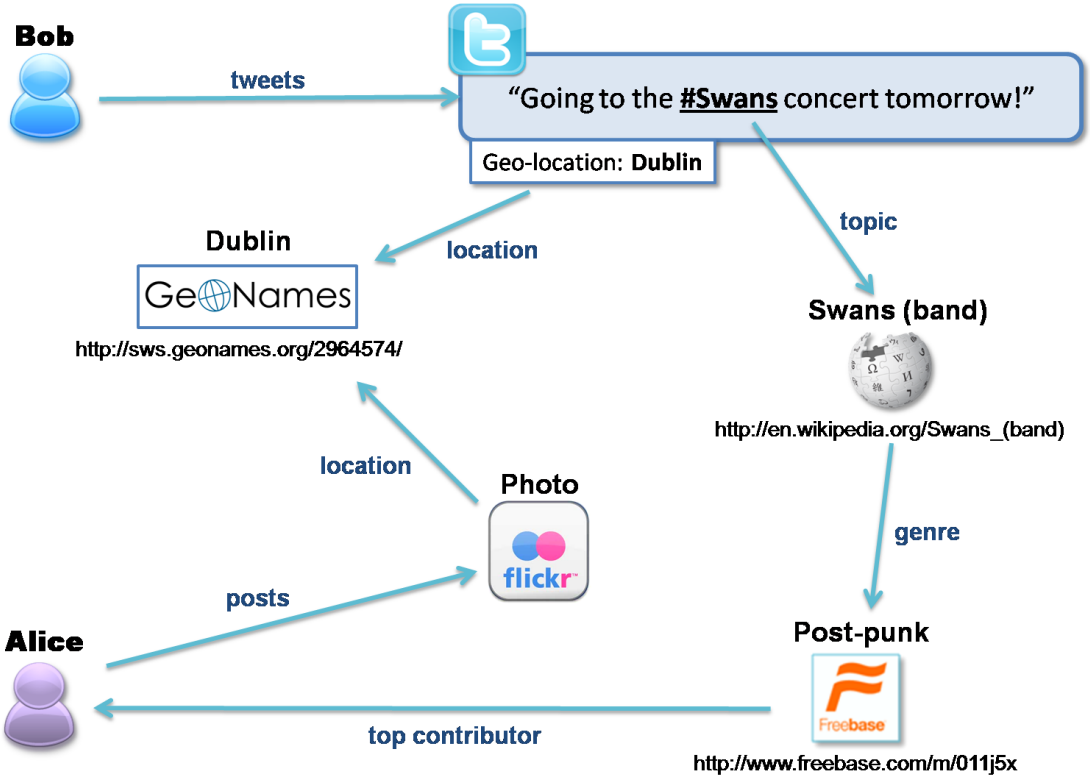

Figure 4.2.: Example of interconnection between Social Web and Web of Data where provenance of data plays a key role.

In Figure 4.2 we illustrate a possible scenario that could ideally happen on the current Web. This example serves as a demonstration of how well connected Social Web and Web of Data could be. This connection is possible thanks to provenance information both on the Social Web and the Web of Data. In the example it is shown how two users could discover to have something in common thanks to social media and semantically interlinked datasets. In particular, Bob posts a status on Twitter related to a music band and the tweet is geo-tagged with Bob's location (Dublin). At the same time another Web user, Alice — who is a very active contributor on Freebase[1] on topics related to a particular music genre ("Post-punk") — posts a photo on Flickr[2], which is also geo-tagged with the same location of Bob's tweet. Therefore, it is clear that Alice and Bob have an interest in common ("Post-punk" music) and are also probably

[1]Freebase is a large collaborative knowledge base consisting of metadata composed mainly by its community members and contributing to the Web of Data with its large public RDF dataset: http://www.freebase.com (accessed January 2014)

[2]Flickr is a popular photo hosting/sharing website: http://www.flickr.com (accessed January 2014)

located in the same area. Hence, Alice could get Bob's tweet as a recommendation from a personalisation system or even the two users could start a conversation and share knowledge about concerts in Dublin. This connection is possible not only because of social media, but also thanks to:

- Linked Data datasets representing knowledge on the Web in a meaningful way (in the example we have Freebase, GeoNames[3] and the semantic export of Flickr [Passant, 2008]);
- Provenance information on the Social Web (e.g. hashtag and user account from Twitter, geo-location from Twitter and Flickr, types of actions and sources, etc.);
- Provenance information on the Web of Data (e.g. metadata about contributors, authorship and versioning on Freebase).

While the example in Figure 4.2 is just an ideal use case that serves as a practical point of view on the problem, on the Web there are many other similar examples. In particular, we distinguish them with the two main directions of exchange of provenance of data:

- **Provenance on the Social Web for the Web of Data.** In this case every social media platform that is exported to the Web of Data thanks to Semantic Web technologies, generates datasets on the LOD Cloud that are the result of activities of their Social Web communities. These datasets should preserve provenance information about their original users and their contributions. Some examples are: Wikipedia and its Web of Data exports DBpedia and Freebase, the export of Last.FM data and users[4], Twitter semantic exports[5], the Facebook Open Graph, etc.
- **Provenance on the Web of Data for the Social Web.** This category includes all the Linked Data datasets that are used by Web applications and services. Provenance information on the datasets fuelling a Social Web application is fundamental when users of the application could analyse the quality of the information they consume. Some examples: the DBpedia, Yago and/or Freebase datasets (which have a social/collaborative nature) are being used by a number of systems, from IBM

[3] http://www.geonames.org (accessed January 2014)
[4] http://datahub.io/dataset/rdfize-lastfm (accessed January 2014)
[5] http://datahub.io/dataset/twarql (accessed January 2014)

Watson [Ferrucci et al., 2010], to Google Knowledge Graph[6]. The BBC publishes semantic datasets and applications about TV/radio programs, music and documentaries [Kobilarov et al., 2009]. The New York Times provides news as Linked Data. Governments such as the USA[7] and UK[8] ones publish open government data, applications and reports to their citizens as results of complex aggregations of various data sources.

The benefits of using data provenance to develop trust on the Web, and the Semantic Web in particular, have been already widely described in the state of the art (see [Li et al., 2010] and [Hartig and Zhao, 2009]). Provenance data provides useful information such as timeliness and authorship of data. It can be used as a ground basis for various applications and use cases such as identifying trust values for pages or pages fragments [Adler et al., 2008], or measuring users' expertise by analysing their contributions [Hoisl et al., 2007] and then personalize trust metrics based on the user profile of a person on a particular topic [Golbeck et al., 2003]. Moreover, providing also provenance meta-data as RDF and making it available on the Web of Data [Hartig, 2009], offers more interchange possibilities and transparency. This would let people link to provenance information from other sources. It provides them the opportunity to compare these sources and choose the most appropriate ones, or the one with higher quality.

Collaborative websites such as Wikipedia have shown the benefit of being able to create and manage very large public knowledge bases[9]. However, one of the most common concerns about these types of information sources is the trustworthiness of their content which can be arbitrarily edited by everyone. The DBpedia project, which aims at converting Wikipedia content into structured knowledge, is then not exempt from this concern. Especially considering that one of the main objectives of DBpedia is to build a dataset such that Semantic Web technologies can be employed against it. Hence this allows not only to formulate sophisticated queries against Wikipedia, but also to link it to other datasets on the Web, or create new applications or mashups [Auer et al., 2007]. Thanks to its large dataset (around 1 billion RDF triples) and its cross-domain nature, DBpedia has become one of the most important and interlinked datasets on the Web of Data [Cyganiak and Jentzsch, 2010] [Bizer et al., 2009]. Therefore ensuring provenance information of DBpedia data is crucial, especially for developers consuming or interlink-

[6] http://www.google.com/insidesearch/features/search/knowledge.html (accessed January 2014)
[7] https://www.data.gov/ (accessed January 2014)
[8] http://data.gov.uk (accessed January 2014)
[9] Statistics about Wikipedia: http://stats.wikimedia.org/EN/Sitemap.htm (accessed January 2014)

ing its content. The same problem applies to many other datasets on the Web, especially for those where the content is collaboratively edited by a community of users (e.g. Freebase, the Google Knowledge Graph, or the Facebook Open Graph[10]), or the published data is the result of complex software extraction and data aggregation processes (e.g. data.gov.uk and other governmental catalogues).

4.2. Provenance on the Social Web for the Web of Data

As described in Chapter 2, the Social Web provides users the ability to select content from across the Web, integrate it, edit it, rate it, publish it, and share it with others. This workflow is similar to the one performed by data journalists, but typically requires very little technical skills, as it is supported by social media platforms such as Facebook or Google+. It is a workflow based on a "consume-select-curate-share" structure and people are not the only actors involved: software agents or programs also play a relevant role [Moreau and Groth, 2013]. This kind of human-computer interaction is often referred to as *social machines*[11] [Berners-Lee and Fischetti, 1999]. In this regard, widely cited examples of large and successful online communities, with thousands of users and software agents collaborating together on a particular project are Wikipedia and Ushahidi[12]. The typical problem with these sources of information is trust on the Social Web. When software bots or malicious users or even just users without expertise can participate in the curation of the online content, it is obvious that quality and reliability of Social Web information is questioned. Hence, in this context provenance is one of the main solutions to the problem. It indicates who contributed to which piece of information, it helps consumers check where information comes from, why it was selected, and how it was edited [Gil et al., 2010].

However, traditional approaches for managing provenance do not address contemporary social media. W3C established a working group to provide recommendations for possible standards (Section 2.2). The work conducted by the W3C Provenance Group provided a core provenance data model to be adopted as a provenance interchange model across heterogeneous systems (see the PROV data model described in Section 2.2.2). The proposed standardisation approach presents still many challenges. In particular it is based on the assumption that there is widespread use of Semantic Web technologies,

[10] https://developers.facebook.com/docs/opengraph/ (accessed January 2014)
[11] http://sociam.org (accessed January 2014)
[12] www.ushahidi.com (accessed January 2014)

or "linked open provenance data" [Hartig, 2009], which is currently not always true for the Social Web.

The only solutions to this problem are either a widespread agreement among the most popular social media sites (Facebook, Twitter, etc.) for the adoption of the standards, or to develop mechanisms for the automated extraction of provenance information and its publication online according to the standards. This way, by showing interesting applications developed on top of provenance, there could be and increase of interest on this important topic. In this direction goes our proposed approach for extracting provenance information from the Social Web. We represent provenance according to the W3C standards and we publish it to the Web of Data providing also an interesting application on top of the exposed data.

The W3C Working Group in [Gil et al., 2010] highlights other challenges related to provenance on social media:

- "No common format and application programmers interface (API) to access and understand provenance information, whether explicitly indicated or implicitly determined.
- Developers rarely include provenance management or publish provenance records.
- No widely accepted architecture solution to managing the scale of provenance records.
- No existing mechanisms for tying identity to objects or provenance traces.
- Incompleteness of provenance records and the potential for errors and inconsistencies in a widely distributed and open setting such as the web."

Partial solution to these challenges is provided in [Barbier et al., 2013] where Barbier et al. describe an approach for reconstructing a network with information propagation, tracking the possible diffusion of information on social media, which is essential for information provenance. "Knowing the provenance of a piece of information published in social media — how the piece of information was modified as it was propagated through social media and how an owner of the piece of information is connected to the transmission of the statement — provides additional context to the piece of information. A social media user can use this context to help assess how much value, trust, and validity should be placed on the information." [Barbier et al., 2013]. Social media can help solving this information provenance problem due to its unique features: user-generated

content, user profiles, user interactions (e.g., links between friends, hyperlinks on blogs, or news articles), and spatial or temporal information. These characteristics can facilitate the reconstruction of a network with information propagation, which is essential for tracking provenance information on the Social Web.

In this work we assume that provenance information is somehow already available in a structured or unstructured form in social media sites. In other words, we do not investigate the aforementioned problem of information propagation discussed by Barbier et al. which is currently an important challenge. Therefore, we use all the information that is already provided as metadata by social media platforms, or that is provided by the websites APIs or can be extracted using parsers and natural language processing techniques. In the following Section 4.2.1 we describe our methodology for: extracting provenance information from social media, representing it using open W3C Web standards, and building useful applications on top of it exploiting the potential of the Web of Data. Although we focus on the particular scenario of Wikipedia and DBpedia, the described methodology can be generalised to other Social Web use cases.

4.2.1. Use Case: Provenance on Wikis

In this section we first overview some related work in the realm of provenance management on the Web of Data and in trust and quality evaluation techniques on wikis. Comparing these two research fields we highlight the limitations that we found in both of them: the former lacks of concrete and well established procedures to support the integration and publication of provenance of non- or semi-structured data on the Web of Data; the latter does not take into account the importance of making the information generated analysing users' edits available as Linked Data and providing details of the steps involved in the analysis. In Section 4.2.1.2, we detail the W7 model for provenance representation, as previously designed by S. Ram et al. [Ram and Liu, 2007], and our implementation of this model with a lightweight ontology built to express it in RDFS. In particular we use the SIOC vocabulary and its extensions since it aims at describing the structure of online communities such as in wikis, and its *Actions* module suits well our need of defining events and user activities in wikis. In Section 4.2.1.3 we also provide an alignment of our model with the Open Provenance Model (OPM), the reference ontology chosen by the W3C Provenance Incubator Group. Then, in Section 4.2.1.4 we describe an application that extracts provenance information from Wikipedia and uses it to provide useful information directly on Wikipedia articles. Our application also

represents provenance using our model, exposing it to the Web of Data and connecting it to DBpedia.

4.2.1.1. Related Work

Research on Wikipedia, and on collaborative websites in general, shows that some information quality aspects (such as currency and formality of language) of Wikipedia are quite high [Lih, 2004]. However, as suggested in [Stvilia et al., 2005], the high quality level of certain aspects of Wikipedia articles does not imply that it is good on other dimensions as well. In fact, a substantial qualitative difference exists in Wikipedia between "featured" articles (high quality articles identified by the community) and normal articles [Stvilia et al., 2005]. For this reason it is important to identify quality measures for Wikipedia articles and estimate the trustworthiness of their content. Then, since the DBpedia content is directly extracted from Wikipedia, the same trust and quality values can be propagated to the DBpedia dataset. However, in order to obtain these values, it is essential to provide detailed provenance information about the data published on the Web.

Another research topic relevant to our work is the evaluation of trust and data quality in wikis. Recent studies proposed several different algorithms for wikis that would automatically calculate users' contributions and evaluate their quantity and quality in order to study the authors' behaviour, produce trust measures of the articles and find experts. WikiTrust [Adler et al., 2008] is a project aimed at measuring the quality of author contributions on Wikipedia. They developed a tool that computes the origin and author of every word on a wiki page, as well as "a measure of text trust that indicates the extent with which text has been revised"[13]. On the same topic other researchers tried to solve the problem of evaluating articles' quality, not only examining quantitatively the users' history [Hoisl et al., 2007], but also using social network analysis techniques [Korfiatis et al., 2006]. Another relevant contribution is in [Demartini, 2007], where the author details the implementation of a system for expert finding in Wikipedia.

From our perspective, there is a need of publishing provenance information as Linked Data from websites hosting a wide source of information (such as Wikipedia) and also from relevant datasets (such as DBpedia). Yet, most of the work on provenance of data is, either not focused on integrating provenance information on the Web of data, or mainly based on provenance for resource descriptions or already structured data. On

[13] http://wikitrust.soe.ucsc.edu/

the other hand, the interesting work done so far on analysing trust and quality on wikis does not take into account the importance of making the analysed data available on the Web of data.

Relevant related research in our context is also presented in [Vrandecic et al., 2010] and [Ceolin et al., 2010]. First, the work by Vrandečić et al. describes a collaborative Web application that allows users to aggregate sources of information on entities of interest from the Web of Data. It takes Wikipedia as its starting point for its entities and it provides the source of every information added by its users. Second, the research presented by Ceolin et al. describes a trust algorithm for event data and an ontology representing events in general, the Simple Event Model. Interestingly the authors provide a discussion of a mapping between OPM and the Simple Event Model using a similar methodology to ours (as we will detail in Section 4.2.1.3).

Overall it is important to mention a similar approach to our work that has been implemented and described in [McGuinness et al., 2006]. The authors propose an algorithm to compute trust values on Wikipedia articles using provenance information extracted from the revision history. The algorithm implemented to compute trustworthiness of assertions is based only on the internal links between articles and more specifically on citations. Hence this work is more focused on computing trust of Wikipedia articles rather than on representing and publishing provenance information to the Web of Data. A vocabulary for annotating the provenance information is used, it is called the Proof Markup Language (PML)[14], but the data used by the experiment has not been published. However, since we focus on representing and publishing provenance of DBpedia to the Linked Open Data, we decided to use popular lightweight ontologies such as SIOC, Dublin Core and ChangeSet[15] to represent edits in Wikipedia and changes to DBpedia statements. These popular ontologies have been integrated and extended with specific modelling solutions to represent more in depth the Wikipedia edits history (for more details see our W7 ontology implementation described in Section 4.2.1.2). Mappings to the OPM ontology have also been provided in order to facilitate the integration with other provenance data, as OPM has been chosen as a reference by the W3C Incubator Group (more details in Section 4.2.1.3). Furthermore with our work we show how we reused existing community ontologies and how these vocabularies can be applied to a concrete use case in order to represent provenance at a triple level and publish it as Linked Data.

[14] http://tw.rpi.edu/portal/Proof_Markup_Language
[15] http://vocab.org/changeset/schema.html

4.2.1.2. Representing Provenance on Wikis Using the W7 Model and RDFS/OWL

The W7 model is an ontological model created to describe the semantics of data provenance [Ram and Liu, 2007]. It is a conceptual model and, to the best of our knowledge, no RDFS/OWL representation of this model has been implemented yet. Hence, in this thesis we focus on an RDFS/OWL implementation of W7 for the specific context of wikis. As a comparison, in their previous work [Ram and Liu, 2009] Ram S. and Liu J. use Wikipedia as an example to theoretically illustrate how their proposed W7 model can capture domain or application specific provenance. Starting from the suggestions and the examples given by these authors we implemented the model described in their publication.

The W7 model is based on the Bunge's Ontology [Bunge, 1977]. In other words, it is built on the concept of tracking the history of the events affecting the status of things during their life cycle. In this particular case we focus on the *data* life cycle. The Bunge's ontology, developed in 1977 by Mario Bunge, is considered as one of the main sources of constructs to semantically model real systems and information systems. While Bunge's work is mainly theoretical, there has been some effort from the scientific community to translate his work into machine readable ontologies [Evermann, 2009][16]. The W7 model can then be seen as an extraction of a part of the constructs described by the Bunge's theories.

The W7 model represents data provenance using seven fundamental elements or interrogative words: *what*, *when*, *where*, *how*, *who*, *which*, and *why*. Hence very similar to the well-known "Five Ws" theory commonly practiced in journalism [Flint, 1917]. All the six interrogative words in the "Five Ws" theory are included in the W7 model. The seventh added word in the W7 model is *which*. In order to generate complete provenance information about a data source, it is necessary to provide an answer to all the seven questions. This model has been purposely built with general and extensible principles, hence it is possible to capture provenance semantics for data in different domains. We refer to [Ram and Liu, 2007] for a detailed description of the mappings between the W7 and Bunge's models, and in Table 4.1 we provide a summary of the W7 elements (as in [Ram and Liu, 2009]).

[16] Evermann J. provides an OWL description of the Bunge's ontology

Provenance element	Construct in Bunge's ontology	Definition
What	Event	An event (i.e. change of state) that happens to data during its life time
How	Action	An action leading to the events. An event may occur, when it is acted upon by another thing, which is often a human or a software agent
When	Time	Time or more accurately the duration of an event
Where	Space	Locations associated with an event
Who	Agent	Agents including persons or organisations involved in an event
Which	Agent	Instruments or software programs used in the event
Why	-	Reasons that explain why an event occurred

Table 4.1.: Definition of the 7 Ws by Ram S. and Liu J.

Having described the structure of the SIOC Actions module in Section 3.3.1.2, and looking at the W7 model summarised in Table 4.1, it is clear why we chose SIOC Actions as core of our model, in particular:

- Most of the concepts in the SIOC Actions module are the same as in the W7 model;
- Wikis are community sites and the Actions module has been implemented to represent dynamic, action-centric views of online communities.

In the following sections we provide a detailed description of how we answered each of these seven questions in order to build provenance data from wikis. Hence, we describe our particular modelling solution which fits our requirements and also integrates well with popular Social Semantic Web vocabularies. However, for the sake of completeness and compatibility, in the next Section 4.2.1.3 we provide mappings between our solution and other standard provenance ontologies.

What The *What* element represents an event that affected data during its life cycle. It is a change of state and the core of the model. In this regard, there are three main events affecting data: *creation*, *modification* and *deletion*. In the context of wikis, each of them can appear: users can (1) *add* new sentences (or characters), (2) *remove* sequences of characters, or (3) *modify* characters by removing and then adding content in the same

position of the article. In addition, in systems like Wikipedia, some other specific events can affect the data on the wiki, for example "quality assessment" or "change in access rights" of an article [Ram and Liu, 2009]; however, they can be expressed with the three broader types defined above.

Since (1) wikis commonly provide a versioning mechanism for their content and (2) every action on a wiki article leads to the generation of a new article revision, the core event describing our *What* element is the creation of an article version. In particular we model this creation, and the related modification of the latest version (*i.e.* the permalink), using the SIOC-Actions model as shown in Listing 4.1.

```
<http://vmuss06.deri.ie/actions#title=Dublin_Core&id=383055>
    sioca:creates <http://en.wikipedia.org/w/index.php?title=Dublin_Core&oldid
        =383055>;
    sioca:modifies <http://en.wikipedia.org/wiki/Dublin_Core>;
    a sioca:Action.
```

Listing 4.1: Representing the "What" element

As we can see from the example above expressed in Turtle syntax, we have a `sioca:Action` identified by the URI `<http://vmuss06.deri.ie/actions#title=Dublin_Core&id=383055>` that leads to the creation of a revision of the main wiki article about "Dublin Core". The creation of a new revision was originated with the modification (`sioca:modifies`) of the main Wikipedia article `<http://en.wikipedia.org/wiki/Dublin_Core>`. Details about the type of event are exposed in the next section about the *How* element, where we identify the type of action involved in the event creation.

How The *How* element in W7 is an equivalent to the *Action* element from Bunge's ontology, and describes the action leading to an event. In wikis, the possible actions leading to an event (*i.e.* the creation of a new revision) are all the edits applied to a specific article revision. By analysing the *diff* between two subsequent revisions of a page, we can identify the type of action involved in the creation of the newer revision. In particular we focus on modelling the following types of edits: *Insertion*, *Update* and *Deletion* of both *Sentence*s and *Reference*s. With the term *Sentence* we refer to every sequence of characters that does not include a reference or a link to another source, and with *Reference* we refer to every action that involves a link or a so-called Wikipedia *reference*. As discussed in [Ram and Liu, 2009], another type of edit would be a *Revert*, or an undo of the effects of one or more edits previously happening. However, in Wikipedia, a revert does not restore a previous version of the article, but creates a new

version with content similar to the one from an earlier selected version. In this regard, we decided to model a revert as all the other edits, and not as a particular pattern. The distinction between a revert and other types of action can be yet identified, with an acceptable level of precision, by looking at the user comment entered when doing the revert, since most users add a related revert comment (the same filtering approach is implemented in [Cosley et al., 2007] with acceptable results)[17].

Going further, and to represent provenance data for the action involved in each wiki edit, we modelled the *diffs* existing between pages. To model the differences calculated between subsequent revisions we created a lightweight *Diff ontology*, inspired by the Changeset vocabulary[18]. Yet, instead of describing changes to RDF statements (which is the scope of Changeset), the Diff model aims at describing changes to plain text documents.[19] This vocabulary does not model differences between any other type of objects such as triples, source code, etc. However, it is designed to be simple and generic enough to model any plain text differences. It has been created together with other ontology engineers and a small community of experts on the Semantic Web domain. Therefore, we followed the recommended practices for creating vocabularies in a collaborative manner ensuring that the result represents a shared view of the modelled domain. Moreover, it has been used and tested in a practical use case (i.e. the wikis use case described here) and it is structurally very similar to the Changeset vocabulary, which is a popular and widely adopted vocabulary describing "diffs" between RDF statements.

The Diff ontology provides a main class, the `diff:Diff` class, with six subclasses: `SentenceUpdate`, `SentenceInsertion`, `SentenceDeletion` and `ReferenceUpdate`, `ReferenceInsertion`, `ReferenceDeletion`, based on the previous *How* patterns.

The main `Diff` class represents all information about the change between two versions of a wiki page (see Fig. 4.3). The `Diff`'s properties `subjectOfChange` and `objectOfChange` point respectively to the version changed by this *diff* and to the newly created version. Details about the time and the creator of the change are provided respectively by `dc:created` and `sioc:has_creator`. Moreover, the comment about the change is provided by the `diff:comment` property with range `rdfs:Literal`. In Figure 4.3 we also display a `Diff` class linking to another `Diff` class. The latter represents one of the six `Diff` subclasses described earlier in this section. Since a single *diff* between two ver-

[17]Note that we could also compare the n-1 and n+1 version of each page to identify if a change is a revert

[18]The Changeset schema: http://purl.org/vocab/changeset/schema

[19]The Diff ontology is publicly available at: http://vocab.deri.ie/diff .

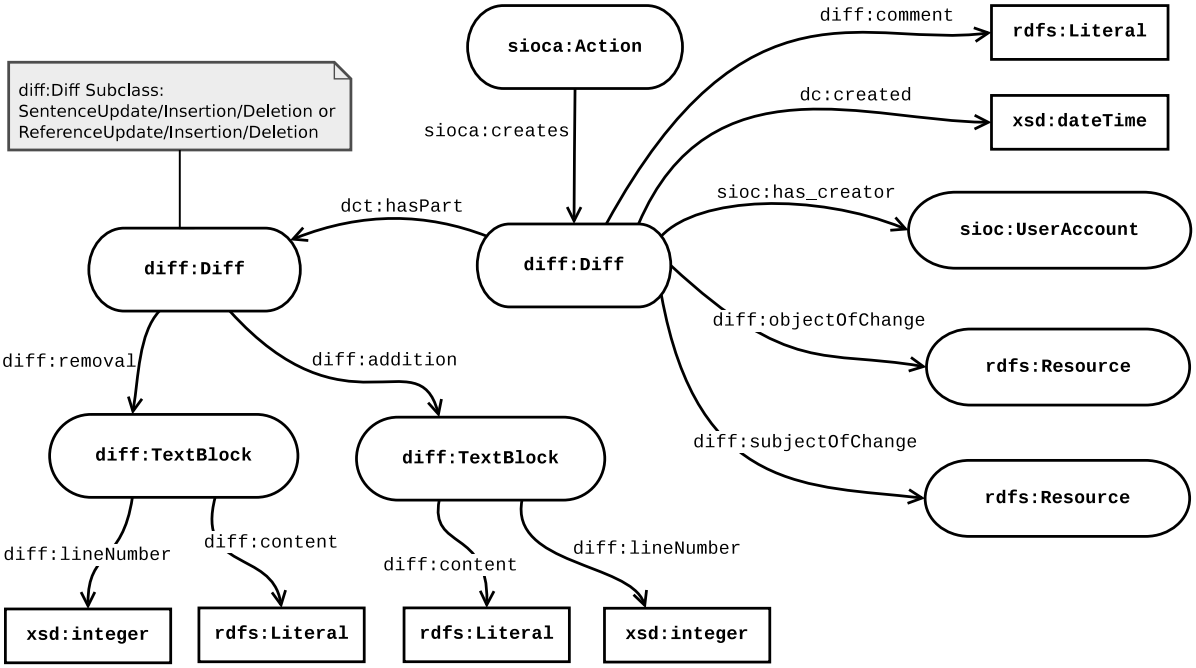

Figure 4.3.: Modeling differences in plain text documents with the *Diff* vocabulary

sions can be composed by several atomic changes (or *"sub-diffs"*), a `Diff` class can then point to several more specific classes (subclasses of `Diff`) using the `dc:hasPart` property. Each `Diff` subclass can have maximum one `TextBlock` removed and one added: if it has both, then the type of change is an *Update*, otherwise the type would be an *Insertion* or a *Deletion*.

The `TextBlock` class is part of the Diff ontology and represents a sequence of characters added or removed in a specific position of a plain text document. It exposes the content itself of this sequence of characters (`content`) and a pointer to its position inside the document (`lineNumber`). It is important to precise that usually the document content is organized in sets of lines, as in wiki articles, but this class is generic enough to be reusable with other types of text organization. To note also that each of the six subclasses of the `Diff` class inherit the properties defined for the parent class, but unfortunately this is not displayed in Figure 4.3 for space reasons.

With the model presented it is possible to address an important requirement for provenance: the reproducibility of a process. Starting from an older revision of a wiki article, just following the *diffs* between the newer revisions and the `TextBlocks` added or removed, it is possible to reconstruct the latest version of the article. This approach goes a step further than just storing the different data versions: it provides details of the entire process involved in the data life cycle.

When The *When* element in W7 is equivalent to the *Time* element from Bunge's ontology, and obviously refers to the time an event occurs, which is recorded in every wiki platform for page edits. As depicted in Figure 4.3, each `Diff` class is linked to the timestamp of the event using the `dc:created` property. The same timestamp is also linked to each `Diff` subclass using the same property (not shown in Fig. 4.3 for space reasons). The time of the event is modelled with more detail in the `Action` element as shown in the following Listing 4.2 [20].

```
<http://example.com/action?title=Dublin_Core#380106133>
   dc:created "2010-08-21T06:36:17Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
   lode:atTime [
     a time:Instant;
     time:inXSDDateTime "2010-08-21T06:36:17Z"^^<http://www.w3.org/2001/XMLSchema#
         dateTime>.
   ];
   a sioca:Action.
```

Listing 4.2: Representing the "When" element in Turtle syntax

In this context we consider actions to be instantaneous. As in [Champin and Passant, 2010] we track the instant that an action is taking effect on a wiki (*i.e.* when a wiki page is saved). Usually, this creation time is represented using `dc:created`. Another option provided by LODE [Shaw et al., 2009] uses the `lode:atTime` property to link to a class representing a time interval or an instant.

Where The *Where* element represents the online "Space" or the location associated with an event. In wikis, and in particular in Wikipedia, this is one of the most controversial elements of the W7 model. If the location of an article update might be considered as the location of the user when updating the content, then this information on the Wikipedia is not completely provided or accurate. Indeed we can extract this information only from the IP address of the anonymous users but not from all the users contributing on the Wikipedia. So at the moment our solution is to just keep track of the IP address of the anonymous users as we can see in SIOC `UserAccount` URIs like this: `http://en.wikipedia.org/wiki/User:96.245.230.136`. We can also link each `UserAccount` with the related IP address using the `sioc:ip_address` property.

Who The *Who* element describes an agent involved in an event, therefore it includes a person or an organization. On a wiki it represents the editor of a page, and it can

[20] For all the namespaces please consult: `http://prefix.cc`

be either a registered user or an anonymous user. A registered user might also have different roles in the Wikipedia site and, on this basis, different permissions are granted to its account. Also, a user account on a wiki can be used by software agents (or bots). Hence, an edit is performed by a user account which is either managed by a person or a software agent. In this work we connect the edits only to the user account and not to the person or bot behind it. We are only interested in keeping track of the user account involved in each event, and not in its role on the wiki. As depicted in Figure 4.4, users are modelled with the `sioc:UserAccount` class and linked to each `sioca:Action`, `sioct:WikiArticle` and `diff:Diff` with the property `sioc:has_creator`. A `sioc:UserAccount` represents a user account, in an online community site, owned by a physical person or a group or an organisation (*i.e.* a `foaf:Agent`). Hence a physical person, represented by a `foaf:Person`, or in general a `foaf:Agent`, could be linked to several `sioc:UserAccount`s.

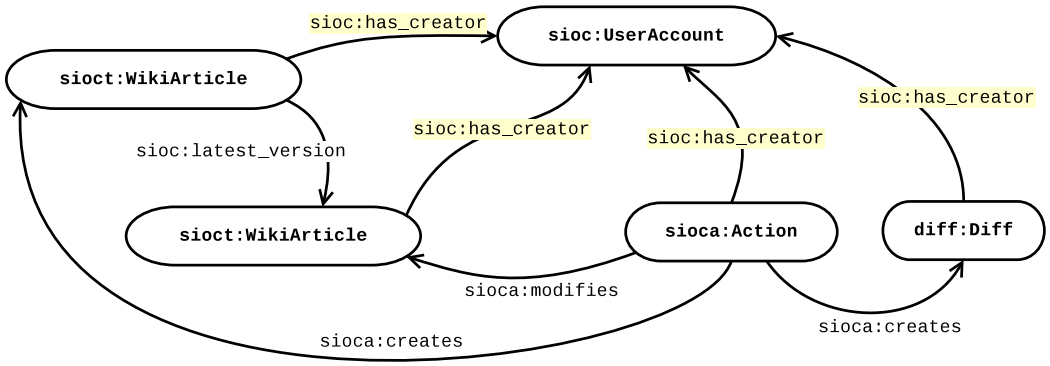

Figure 4.4.: Modeling the *Who* element with `sioc:UserAccount`

Which The *Which* element represents the programs or the instruments used in the event. In our particular case it is the software used in editing the event, which might be a bot or the wiki software used by the editor. Since there is not a direct and precise way to identify whether the edit has been made by an human or a bot, our model does not differentiate that. A naive method could be to look at the username and check if it contains the "bot" string.

Why The *Why* element represents the reasons behind the event occurrence. On Wikipedia it is defined by the justifications for a change inserted by a user in the "comment" field. This is not a mandatory field for the user when editing a wiki page but the Wikipedia guidelines recommend to fill-in this text field. We model the comment left

by the user with a property `diff:comment` linking the `diff:Diff` class to the related `rdfs:Literal`.

4.2.1.3. Alignment With the Open Provenance Model (OPM) and the PROV Ontology

Our proposed modelling solution is a particular implementation specific to the context of wikis. It is important to note that several generic ontologies representing provenance information have been developed. The scope of these vocabularies is to provide general purpose structures and terminologies that describe provenance information across different sets of application domains. Depending on each specific domain, it is then possible to refine and integrate these generic models with more specific vocabularies. The benefits of using common popular ontologies for provenance are clearly the interoperability of the applications using and producing provenance data, and the easy exchange of data between different sources and domains.

The W3C Provenance Incubator Group (see Section 2.2.2) published a document containing mappings between the most relevant provenance ontologies[21]. In this document the ontology taken as reference for the mappings is the Open Provenance Model (OPM) [Moreau et al., 2009]. Continuing that effort, in April 2013, the W3C Provenance Working Group published as a W3C Recommendation a new standard ontology for representing provenance: PROV-O, the PROV ontology (as described in Section 2.2.2). In this section we provide mappings for our modelling solution to both the original OPM model and the new PROV ontology. As the core concepts of the PROV ontology follow the same structure (and mostly also the naming convention) of the OPM ontology, we first provide a comparison of our model with the OPM one and then we just describe the few differences between OPM and PROV in our case.

It is important to note that our "SIOC-based" modelling solution for provenance on wikis is lightweight and can be generalised to other Social Web use cases. Our solution is mainly focused on the reuse of existing terms from popular Semantic Web vocabularies (i.e. SIOC, FOAF, DC and our own Diff vocabulary). This facilitates the integration with existing Social Web datasets on the LOD cloud. However, the W3C Provenance Working Group recommended, at first, the adoption of the OPM model for provenance of data on the Web. Then, as already mentioned above, OPM evolved into the PROV

[21] The document is available at: http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings (accessed January 2014).

ontology and became the new recommended W3C standard. The PROV ontology (and OPM) has been designed to be extended and refined by ontology engineers according the the particular use case. It provides just the main structure of a vocabulary for modelling provenance. Therefore, we consider our "SIOC-based" model as an extension of the PROV (and OPM) model and here we provide mappings between the different solutions. This shows that our model is compliant with the recommended standards.

OPM describes data life cycles in terms of *processes* (events or "things" happening), *artifacts* ("things" involved in a process), and *agents* (entities controlling "things" happening). These three are kinds of nodes within a graph, where each edge denotes a causal relationship. Edges have named types depending on the kinds of node they relate:

- a process *used* an artifact;
- an artifact *was generated by* a process;
- an artifact *was derived from* another artifact;
- a process *was triggered by* another process;
- a process *was controlled by* an agent.

As described in the W3C document providing the mappings, the motivations for the choice of the OPM (and PROV) are: (I) it is a general and broad model that encompasses many aspects of provenance; (II) it already represents a community effort that spans several years and is still ongoing, already benefiting from many discussions, practical use, and several versions; (III) many groups are already undergoing efforts to map their vocabularies to OPM or PROV. For these reasons, and in order to align to the W3C Incubator Group's choice, we defined the ontology mappings between OPM, PROV and our proposed model. Hence here we follow the same procedures used by the W3C Group.

The mappings, summarised in Table 4.2, are expressed using the SKOS vocabulary [Miles and Bechhofer, 2009]. The SKOS mapping properties are `closeMatch`, `exactMatch`, `broadMatch`, `narrowMatch` and `relatedMatch`. These properties are used to state mapping (alignment) links between SKOS concepts in different concept schemes, where the links are inherent in the meaning of the linked concepts. In the table we also provide a column with RDFS alignment properties. By using RDFS for mappings we benefit of reasoning capabilities over the data in our triplestore, hence our local RDF

Terms from our "SIOC-based" model (subject)	SKOS Mappings	RDFS Mappings	Terms from Reference Model (OPM/PROV) (object)
sioca:Action	skos:broadMatch	rdfs:subClassOf	opm:Process / prov:Activity
sioca:DigitalArtifact, sioct:WikiArticle, diff:Diff	skos:broadMatch, skos:broadMatch, skos:broadMatch	rdfs:subClassOf, rdfs:subClassOf, rdfs:subClassOf	opm:Artifact / prov:Entity
sioc:UserAccount	skos:relatedMatch	—	opm:Agent / prov:Agent
sioc:previous_version	skos:broadMatch	rdfs:subPropertyOf	opm:wasDerivedFrom/ prov:wasDerivedFrom
sioca:uses, sioca:modifies	skos:broadMatch, skos:broadMatch	rdfs:subPropertyOf, rdfs:subPropertyOf	opm:used / prov:used
(sioca:creates)	—	—	opm:wasGeneratedBy/ prov:wasGeneratedBy
sioc:has_creator	skos:relatedMatch	—	opm:wasControlledBy/ prov:wasControlledBy

Table 4.2.: Mappings between Open Provenance Model/PROV and our proposed model based on SIOC terms.

store can be queried using OPM-based queries (assuming that RDFS inference support is available in the store).

To better understand the defined mappings and the reasons behind our choices we refer to the diagram displayed in Figure 4.5. In the diagram we show an implementation of the two models under comparison in this section. The one on the top represents our proposed "SIOC-based" model while the other one on the bottom the OPM. To note that the same instances, represented with different classes between the two models, are depicted with the same colours. Moreover, some properties not strictly relevant in this context have been omitted for more clarity, only the terms under comparison between the two models are displayed. In the following part of this section more details about this diagram are provided.

As summarised in Table 4.2, the first analysed mapping is about the `opm:Process` class which represents one or more actions "performed on or caused by artifacts, and resulting in new artifacts. On the other hand the `sioca:Action` is a timestamped event involving a user and a number of digital artifacts. Therefore we can define the *Action* class as more specific (narrower) than the *Process* one, since it is limited to a timestamped instant and to digital artifacts. As regards the artifacts indeed, in the OPM model they are defined as "immutable pieces of state, which may have a physical embodiment in a physical object, or a digital representation in a computer". While in the SIOC Actions module only the concept of *Digital Artifact* is contemplated. Even though the definition of `sioca:DigitalArtifact` is broad and generic (*i.e.* "Anything that can be the object of an Action"), we see this concept as narrower than the OPM one because

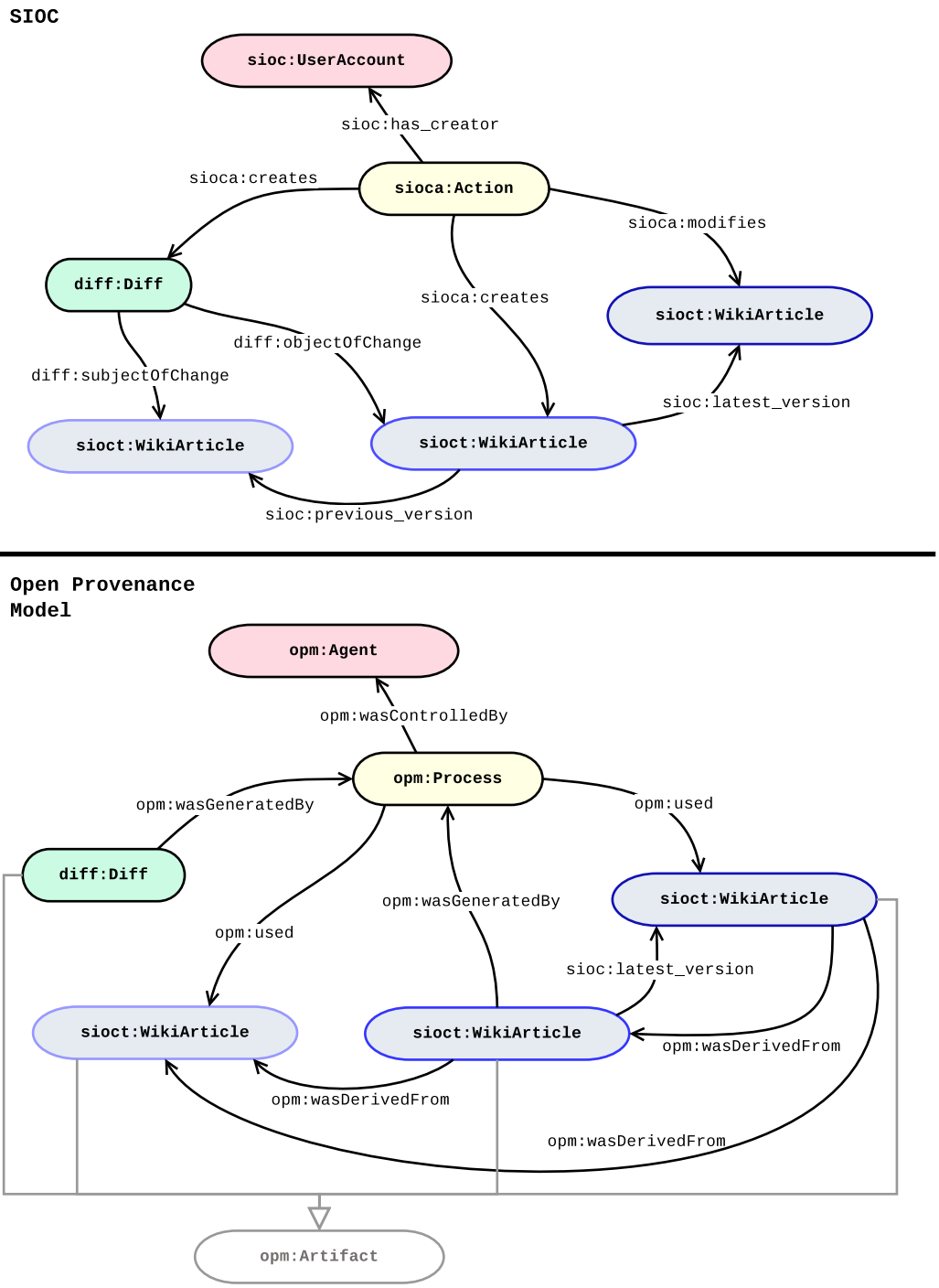

Figure 4.5.: Comparison between our proposed modelling solution using SIOC (and its modules) and a solution using the Open Provenance Model (OPM). The same entities modelled with different classes are identified with the same colour.

it is restricted to *digital* objects. To the list of the artifacts we also included other objects like `sioct:WikiArticle` and `diff:Diff`. These are the artifacts involved in our context of wikis, and obviously they are defined as narrower concepts of the `opm:Artifact` class. In Figure 4.5 the aforementioned artifacts are defined as subclasses of the `opm:Artifact` class.

An important element of the provenance dynamics is the `Agent` or the entity "acting as a catalyst of a process, enabling, facilitating, controlling, or acting its execution" (as defined in OPM). The agent in our case is the user that contributes to the data on the wiki through his/her user account. The `sioc:UserAccount` class is defined as the representation of the account with which the user created the `Action`. Hence this concept is only related to the *Agent* concept since the user and his/her account are two disjoint concepts. For the same reason the properties `opm:wasControlledBy` and `sioc:has_creator`, which link a *process* or an *action* to an *agent* or a *user account*, have a `skos:relatedMatch` assigned.

In OPM five causal relationships (also called *arcs* or *edges*) are recognised. The `wasDerivedFrom` property links an artifact to another artifact that was a cause of its existence. As regards the data in wikis we have the mechanism of different versions of the data that are sequentially created one after the other. Hence the SIOC properties interlinking subsequent revisions (`previous_version` and related `next/latest_version`) have the same causal meaning, but limited to a more specific context. The arc `opm:used` defines the relation between a *Process* and an *Artifact* that has been necessary in the completion of the process itself. The *Process* requires the existence of the *artifact* to initiate/terminate. Two properties in the Actions module are related to this property: `sioca:uses` and `sioca:modifies`; the latter is a sub-property of the former which points to "a digital artifact involved by the action, existing before and after it". Since in the two models the existential requirement is persistent, the SIOC term is narrower than the OPM one because of its limitation to digital artifacts. As regards the `sioca:modifies` property, its definition is: "a digital artifact significantly altered by the action"; in our case this property is used to link the *Action* to the latest version of a wiki article, the one with an alias name that does not change over the time (*e.g.* `http://en.wikipedia.org/wiki/Ireland`). On the other hand, each single revision[22] is "created" (property `sioca:creates`) by the *Action*. This situation is closely matched by the OPM term `wasGeneratedBy`, but this does not have an alignment with the SIOCA

[22]Each revision in Wikipedia has a URI that identifies the ID of the version, *e.g.:* `http://en.wikipedia.org/w/index.php?title=Ireland&oldid=384683529`

term `creates` because they can be considered as *inverse* properties. To clarify, looking at the diagram in Figure 4.5, the three `sioct:WikiArticle` objects are (from the left to the right): the older modified revision of an article, the newer revision, and the latest alias version of the article that does not change URI.

As regards the new PROV W3C Recommendation, as anticipated in this section, its ontology is very similar to the OPM model. In fact, the OPM ontology has been taken as basis for the development of the PROV ontology by the W3C Provenance Working Group. We already provided an overview of PROV in Section 2.2.2 and in Figure 4.6 we show a comparison between the core concepts and predicates of both the ontologies. As we can see from this diagram the two vocabularies are almost identical in their core terms and a mapping of our modelling solution against the PROV terms is straightforward, as summarised in Table 4.2.

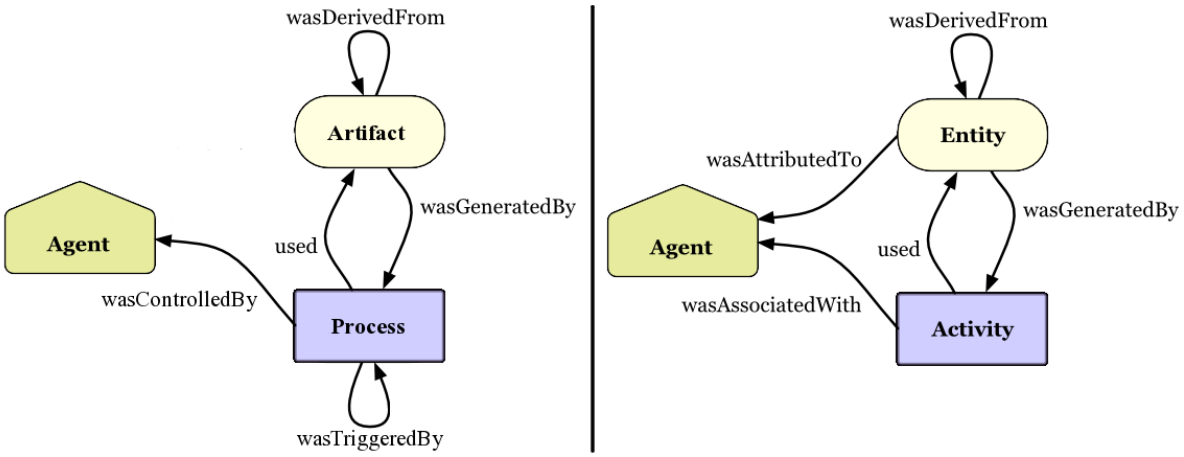

Figure 4.6.: Comparison between the core elements of the OPM (on the left) and PROV (on the right) ontologies

4.2.1.4. Application Using Provenance Data on Wikipedia

Collecting the Data from the Web The first step consists in collecting Wikipedia edits and building related diffs, as well as translating them into RDF. This information is used at a later stage to compute the provenance information, both in Wikipedia and DBpedia. To do so, we designed a script in order to get these information not only for a single page, but for a whole set of pages, belonging to the same category. Practically, the script:

- executes a SPARQL query on the DBpedia endpoint to get the subcategories of the seed one;
- stores these categories (hierarchically represented with SKOS) in a local triplestore;
- queries the DBpedia endpoint to identify all articles belonging to any of theses categories;
- generates (and stores locally) RDF data for each article using the SIOC-MediaWiki exporter that we previously described (Section 3.4.3.1);
- for each article it looks recursively for the previous versions and exports them in RDF.

Figure 4.7 describes the above steps involved in the whole provenance data collection process. Identifying pages in the same category can hardly be done using only Wikipedia, and using DBpedia here (in combination with the former) provides a clear advantage.

Based on this dataset, a second script calculates and models the diff between all consecutive versions of the articles using the Wikipedia API. The API provides us HTML pages with the diff between two revisions, we need to parse these pages and then create the `Diff` objects modelled with the Diff vocabulary described previously in Section 4.2.1.2. Information about the editor, the timestamp, the comment and the ID of the versions collected at the previous step are merged with the *diffs* objects generated in this step. The script also identifies the type of change that happens between versions. This is done by comparing two consecutive versions to identify if the change was an *Insertion* or an *Update* or a *Deletion*. Then, we identify if the change involved a reference or a normal sentence by parsing the content of the `TextBlocks` inside each `Diff`. That way, our export models changes not only as `diff:Diff` instances, but more precisely as *Sentence* or *Reference Insertion/Update/Deletion*. As for the previous extraction, all RDF information about the diffs is stored in the local triple-store, which contains all versioning and *diff* information about pages, modelled using SIOC, SIOC Types, SIOC Actions and the Diff vocabulary. Also, based on the mappings that we defined with OPM, this local store can be queried using OPM-based queries, providing that RDFS inference support is available in the store.

To evaluate this first step, we collected two datasets:

- a first one collecting all articles under the *"Semantic Web"* Wikipedia category (on the English Wikipedia) and all its subcategories.

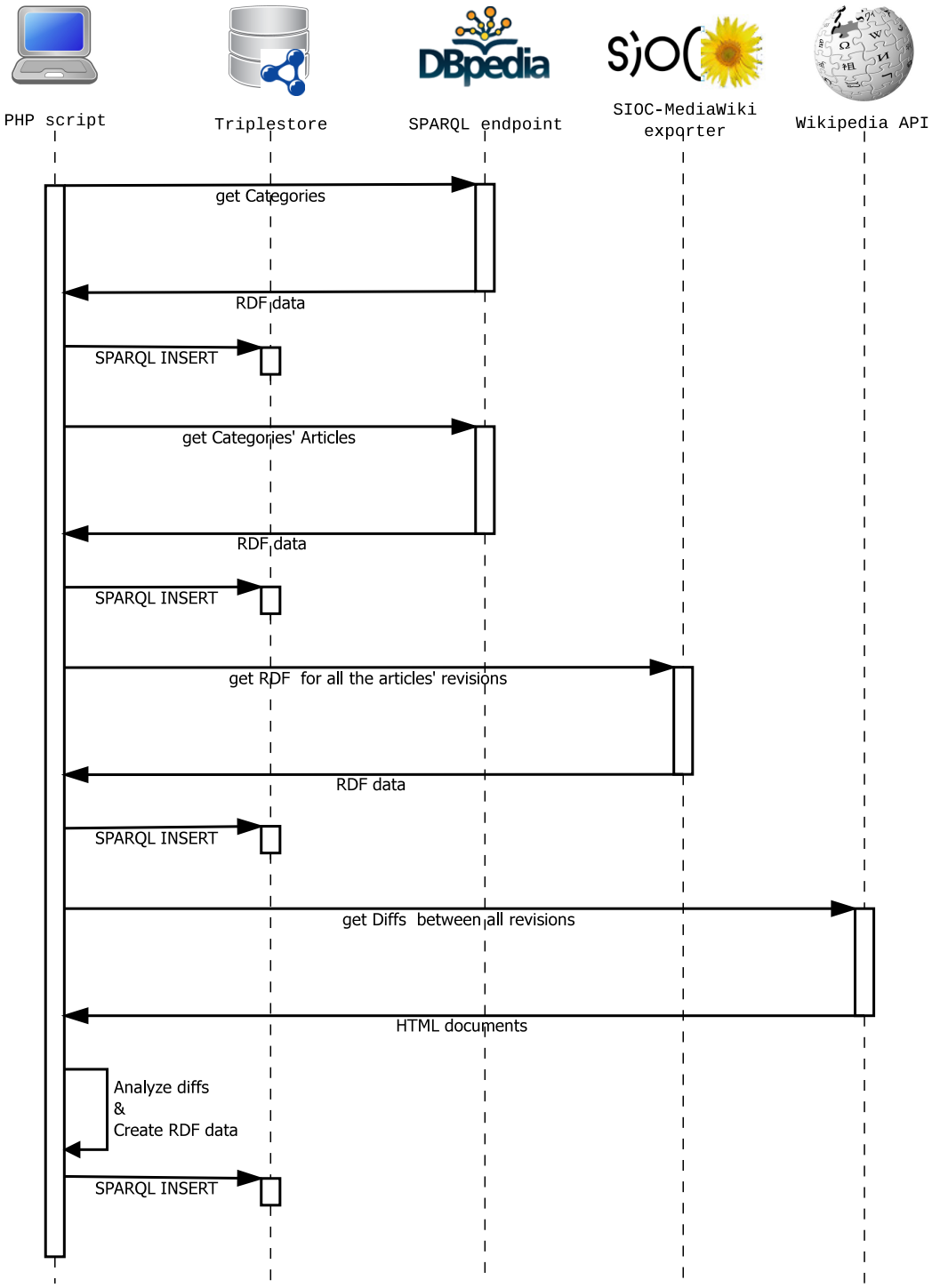

Figure 4.7.: Activity diagram of the provenance data extraction framework

- another one collecting all articles belonging to both *"World Heritage Sites in Italy"* and *"Cities and towns in Emilia-Romagna"* categories.

For the second one, we considered the intersection of the two groups of articles and consequently identified articles about World Heritage Sites in the Italian region Emilia-Romagna. Once again, this particular information cannot be directly retrieved from the Wikipedia articles, as the category does not exist, and has been obtained using a simple SPARQL query on DBpedia.

We also ran the diff extraction algorithm for the *"Semantic Web"* category. It generated data for all the 126 wiki articles belonging to this category and its subcategories recursively (9 categories in total). The total number of triples in the local triplestore for the "Semantic Web" use case is almost 1.5 million triples, for a total of 8656 revisions.

A Firefox Plug-in for Provenance on Wikipedia While our script collects and extracts information from Wikipedia, it is only of limited interest in its current form. The second layer of our framework thus aims at making this information available on the Web (1) directly through Wikipedia pages and (2) both for humans and machines. It thus can be used by people browsing Wikipedia — that directly want to get an overview of the page (or the category) contributing users — or by agents that want to get statistics about these pages in a completely automated manner. The data stored in our triplestore is publicly available on the Web and accessible to software applications as RDF data directly using a RESTful Web service[23]. The other part of our application that aims at making our data more accessible to humans, is also based on the previous triplestore. It consists in a Greasemonkey script[24], which identifies the Wikipedia page currently browsed and sends this to a PHP script, which returns information about the page, using SPARQL queries run on the triplestore. This information is made available on the top of each Wikipedia article, and exposes information about the most active users on the article and their edits. In addition, as we will see next, this application also provides links to RDF representation of this information available through our Web service. By being a Greasemonkey script, it can be installed by anyone on Mozilla Firefox browsers as well as other popular Web browsers supporting it. This also imply that this information is not restricted to RDF-savvy users (as if being in the RDF store only), but can simply be browsed in the standard Wikipedia.

For each page, the script identifies the top contributors (identified as the ones that made the most edits), and computes for each of them:

[23]The Web service that provides raw RDF data is available at: http://vmuss06.deri.ie/WikiProvenance/index.php

[24]http://www.greasespot.net/ (accessed January 2014)

- the total number of edits;
- the percentage of "ownership" on the page (*i.e.* the percentage of their edits compared to all the edits of the article);
- the number of lines added;
- the number of lines removed.
- the number of lines added or removed on all the articles belonging to the category *"Semantic Web"*.

This information is then available as a table on the top of the page, as seen in Figure 4.8 (top figure) for the "Linked Data" page. For categories, similar information is identified, albeit identifying these statistics for all pages of the category, and not for a single page. Browsing a wiki category page, the application shows a list of the users with the biggest number of edits on the articles of the whole category (and related subcategories). Additionally, it displays the related percentages of their edits compared to the total edits on the category. It also exposes a list of the most edited articles in the category during the last three months. A screenshot of the result for categories can be seen in Figure 4.8 (bottom). We can see, at the bottom of each table a link pointing to a page where a longer list of results will be displayed.

Furthermore, to make that information available to machines, these statistics are made available in RDF. We especially relied on SCOVO[25], the Statistical Core Vocabulary [Hausenblas et al., 2009]. The decision to use this vocabulary has been made at the time of the design and implementation of our framework. At that time, SCOVO was the most complete and popular vocabulary for modelling statistical information, however it has been recently superseded by a W3C Recommendation with a new vocabulary called RDF Data Cube[26]. This new vocabulary would be suitable for our use case but the lightweight nature of SCOVO adapts well to our simple use case. SCOVO relies on the concepts of `Item`s and `Dimension`s to represent statistical information. In our context, the `Item` is one piece of statistical information (*e.g. user "X" edited 10 lines on page "Y"*) as displayed in the example in Listing 4.4. In a description of an `Item` various `dimension`s are involved:

[25] http://vocab.deri.ie/scovo (accessed January 2014)
[26] http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/ - RDF Data Cube, W3C Recommendation, 16 January 2014

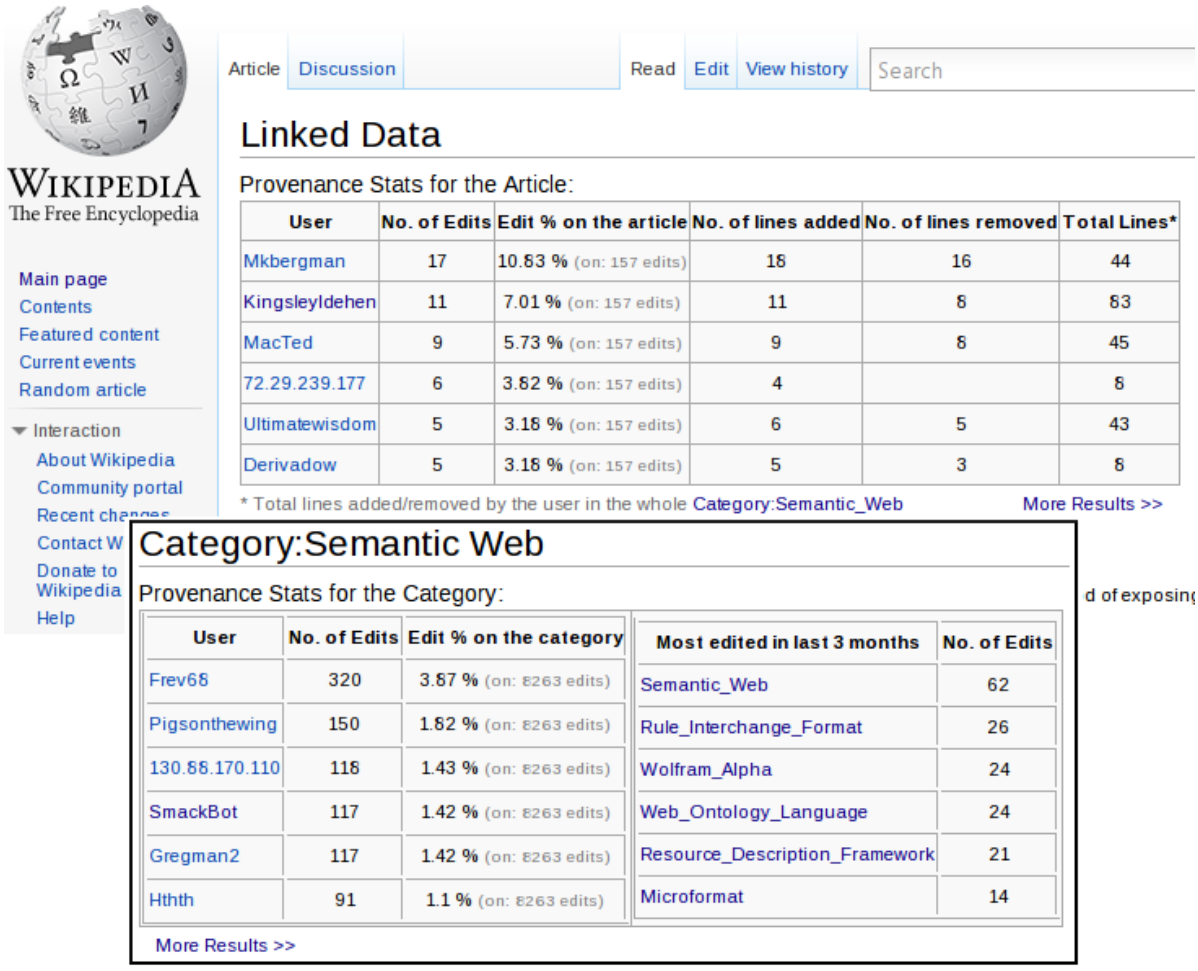

Figure 4.8.: A screenshot of the application on the "Linked_Data" page and the table from the Category "Semantic_Web" page

- the type of information that we want to represent (number of edits, percentage, lines added and removed etc.);
- the URI of the page or the category impacted;
- the URI of the user involved.

Hence, we created four instances of `scv:Dimension` to represent the first dimension (as in Listing 4.3), and relied then simply on the `scv:dimension` property for the other ones. One issue yet with this approach is that it does not differentiate the dimension related to the user and the one related to the page, which is a limitation of SCOVO itself[27]. In the future we may either create new properties, or check other recommended vocabularies

[27] We considered using `sioc:has_creator` but semantically is not exactly the same, as the user is not creating the scovo:Item *per se*, but is just a part of its statistical information.

for representing statistics on the Web of Data such as the RDF Data Cube vocabulary[28], and SDMX[29]. As an example, Listing 4.4 represents that the user `KingsleyIdehen` made 11 edits on the *SIOC* page.

```
@prefix st:<http://vmuss06.deri.ie/stats#>
@prefix scv:<http://purl.org/NET/scovo#>

st:WikiEdits rdfs:subClassOf scv:Dimension ;
  dc:title "Edits in wikis" .

st:Edits a :WikiEdits ;
  dc:title "Number of edits" .
st:EditsPercentage a :WikiEdits ;
  dc:title "Percentage of the overall number of edits" .
st:LinesAdded a :WikiEdits ;
  dc:title "Number of lines added" .
st:LinesRemoved a :WikiEdits ;
  dc:title "Number of lines removed" .
st:LinesInCategory a :WikiEdits ;
  dc:title "Number of lines added or removed on the category" .
```

Listing 4.3: Representing a SCOVO Dimension for the number of edits in wikis

```
@prefix st:<http://vmuss06.deri.ie/stats#>
@prefix scv:<http://purl.org/NET/scovo#>

st:title=SIOC&user=KingsleyIdehen a scovo:Item ;
   rdf:value 11 ;
   scv:dimension st:Edits ;
   scv:dimension <http://wikipedia.org/wiki/SIOC>;
   scv:dimension <http://wikipedia.org/wiki/User:KingsleyIdehen>.
```

Listing 4.4: Representing the number of edits by a user with SCOVO

With this single script, one can get the same information displayed using the Greasemonkey script and also to have the raw RDF description of the page requested. These scripts (the extraction framework and the provenance visualisers) are available at http://vmuss06.deri.ie/WikiProvenance/index.php, as well as the browser plug-ins. Also, a short video demonstrating the application is available at the address http://vmuss06.deri.ie/WikiProvenance/video/.

[28] http://www.w3.org/TR/vocab-data-cube/ - W3C Candidate Recommendation 25 June 2013
[29] Statistical Data and Metadata eXchange: http://sdmx.org/ (accessed January 2014)

4.3. Provenance on the Web of Data for the Social Web

Not only provenance of data on the Social Web is useful for its own users but also provenance on the Linked Open Data is essential for several purposes. By providing provenance meta-data as RDF and making it available on the Web of Data, more transparency and interchange possibilities are offered [Hartig, 2009]. This would let people link to provenance information from other sources. It provides them the opportunity to compare these sources and choose the most appropriate one or the one with higher quality. It also allows for Semantic Web developers to be in control of the origins and quality of the data used for their applications. Finally, it supports Social Web users in a transparent way by providing them with novel advanced applications employing semantic technologies and heterogeneous provenance-aware datasets. This is the effect of the positive feedback loop generated by provenance on the Web, as depicted in Figure 4.1.

As an example, in the following Section 4.3.1, we describe the close connection between a social site such as Wikipedia and the Web of Data equivalent DBpedia. This is a clear example of how Social Web and Linked Data can be interconnected and mutually supporting each other. In this context we can consider, for instance, the following *scenario*.

Wikipedia users perform many different types of edits on a diverse set of articles. In order to better understand users contributions, or to profile their interests or expertise, it is necessary to analyse their edits by aggregating and processing their activities and edits history. Wikipedia edits are quite diverse as regards their extension, type of edit and content. Some edits even propagate to the Web of Data as they modify the structured information about Wikipedia articles recorded on DBpedia. By recording all this information about user edits, interlinking it using standard semantic technologies and recording their provenance both on the Social Web side and the Web of Data side, novel applications and opportunities can be developed. For example, it would be possible to provide users with an application that aggregates, understands and describes users interests, expertise and level of contribution on Wikipedia by interlinking all this information. Moreover, this application could share and interchange data with many other Social Semantic Web applications without problems of integration of data models and so on.

Having this scenario in mind, and motivated by the goal of providing a methodology for profiling user interests on the Social Semantic Web, we focused our attention on

provenance of data as fundamental basis for a complete user profiling methodology (more details in Section 4.4). Hence, to integrate data provenance with social data and the Web of Data, semantic representation of provenance on both the Social Web and the Web of Data is crucial.

In particular in this chapter, as an experiment evaluating the validity of the methodology, our work focused on delivering provenance information about DBpedia statements. Associating provenance information to each one of the million triples in DBpedia could be relevant in several use cases, especially for applications built on top of it. For example, by indicating by whom and when a triple was created (or contributed by), it could let any application flag, reject or approve this statement based on particular criteria. A site could decide to reject statements considered as being too new (so not having been checked by the page editor and the community), or because the author is not trusted in the area (*e.g.* the domain or range of the statement).

This need for provenance management in DBpedia is even more relevant in the case of the "Live" version of DBpedia [Hellmann et al., 2009] and the introduction of a new provenance element in the N-Quads DBpedia dump. This last feature is available only by downloading the N-Quads version of the DBpedia dump and it includes a provenance URI to each statement. The provenance URI denotes the origin of the extracted triple in Wikipedia by exposing the line and the section of a Wikipedia article where the statement has been extracted from. This is a first promising step that demonstrates the growth of interest in the topic. On the other hand with DBpedia Live, since information from Wikipedia will be immediately available in RDF and may be injected live in third party applications, it is important to provide this applications with means to decide if they should accept a statement or not. Finally, more than trustworthiness, provenance in DBpedia can be used for other purposes such as expert finding or social network analysis, focusing on the object-centred sociality vision, by identifying people contributing and socializing around similar resources. In both cases, more than on resources, we could rely on categories, that can be identified by selecting all resources associated to a particular DBpedia category, or more completely through SPARQL queries, such as identifying which people are contributing to pages about Web standards contributed by a particular organization.

To provide such features, we built a framework that

- on the one hand, extracts provenance information for DBpedia, using Wikipedia edits and

- on the other hand, makes that information available on the Web of Data, so that it can be used when building applications based on DBpedia.

We thus propose a twofold approach for provenance management *from and for* the Web of Data, combining Social Web paradigms (editing behaviours in Wikipedia) and Linked Data (provenance information about DBpedia in RDF). The system also makes Wikipedia edits available in RDF, letting Web Scientists interested in Wikipedia collaboration patterns get relevant data using Semantic Web techniques and tools, rather than learn the Wikipedia API.

4.3.1. Use Case: Provenance on DBpedia

In this section we describe a framework to track provenance about DBpedia resources and statements, based on Wikipedia provenance information. Moreover, an application that uses provenance on DBpedia and exposes it in a meaningful way to users will be detailed. With this particular use case we demonstrate how the management of provenance on the Web of Data can be directly dependent on the provenance on the Social Web and vice-versa. We show how provenance can be useful not only for the development of the Linked Data initiative but also for Social Web users. In particular, our goal is not only to provide provenance data from Wikipedia, but also to keep track of the changes happened in Wikipedia and to identify what are the effects of these changes on the DBpedia dataset. In this section we show how we identify the authors of the triples stored in the DBpedia dataset and how we can relate them to the provenance details previously generated from the corresponding Wikipedia articles. The built application leverages the provenance data created for Wikipedia and combines it with the DBpedia extraction procedures. In order to retrieve the set of properties mapped from the infobox properties on Wikipedia to DBpedia, we took the mappings defined on the related DBpedia wiki[30]. In this wiki it is possible to find the infobox-to-ontology and the table-to-ontology mappings which are used by the DBpedia extraction framework. The framework collects the templates defined in the wiki and extracts the Wikipedia content according to them.

As described in Section 4.2.1.4 for our specific use case about the "World Heritage Sites in Emilia-Romagna" we collected pages belonging to two different categories. All the articles resulting from the intersection of the two categories use one particular Wikipedia Infobox called "Infobox Italian comune". This "table template" defines the

[30] http://mappings.dbpedia.org/index.php/Main_Page (accessed January 2014)

properties associated with all the articles about cities in Italy. The structure of this template is shown in Listing 4.5, where part of the Infobox source text of the article "Modena" is displayed. The *wiki text* displayed is then translated and rendered by Wikipedia in a table usually on the top right corner of the page.

```
{{Infobox Italian comune
| name  = Modena
| official_name = Comune di Modena
| native_name =
[...]
| postal_code = 41100
| area_code = 059
| website = {{official|http://www.comune.modena.it}}
| footnotes =
}}
```

Listing 4.5: An excerpt of the Wikipedia "Infobox Italian comune" from the article "Modena"

Once the mappings between Wikipedia and DBpedia were retrieved and the provenance data for the Wikipedia articles generated and stored using our data extraction framework, our application was ready to be implemented. A PHP script has been developed to analyse the content of the `TextBlock`s of each `Diff` stored in our dataset (Section 4.2.1.2). A single SPARQL query is necessary to get the content of the diffs which are probably related to some changes happened in the Infobox part of the wiki article. The aforementioned query is displayed in Listing 4.6. For each change happened in the first 30 lines of the article's revisions it returns the user, the timestamp, the object of change, the content of the line changed and the position of the line in the article. The reason for the line number restriction is because, in our case, the Infobox properties are always positioned in this part of the articles.

```
SELECT distinct ?user ?date ?obj ?content ?line WHERE {
  GRAPH <cities> {
  ?diff rdf:type diff:Diff ;
    dct:hasPart ?subdiff ;
    dc:created ?date ;
    sioc:has_creator ?user ;
    diff:objectOfChange ?obj . FILTER regex(?obj, ".$pagetitle.").
  ?subdiff ?addorrem ?txtblk .
  ?txtblk rdf:type diff:TextBlock ;
    diff:content ?content ;
    diff:lineNumber ?line . FILTER (?line < 30).
  }
}
```

Listing 4.6: A SPARQL query to retrieve the lines changed between all the revisions of an article. Line numbers should be less than 30.

Also note that in Listing 4.6 the title of the article is represented by the PHP variable `$pagetitle`.

The application then analyses each line content returned by the query to identify the changes that actually involved the Infobox properties. For each of the changes matching the requirements, their details (user, timestamp, page version, etc.) and the related DBpedia property affected by the change, are stored again in the local triplestore. The results are semantically modelled using the SIOC Actions-based model previously described in Section 4.2.1.2. The only difference here is the use of the Changeset vocabulary[31] to model the changes of the DBpedia triples caused by the Wikipedia Infobox modifications. As described in Section 4.2.1.2 the Changeset protocol [Ltd., 2011] is similar to the Diff model we adopted in this work. Instead of having a `Diff` class that points to added or removed `TextBlocks`, the Changeset vocabulary defines a `cs:ChangeSet` class that points to the *resources* subject and object of change and to the `rdf:Statements` added and removed. Each `Statement` is then composed by one `rdf:subject`, one `rdf:predicate` and one `rdf:object`. Similarly to what previously described, a `sioca:Action` is then linked to a `cs:ChangeSet` with the property `sioca:creates`. In Listing 4.7 we show a modelling example of a ChangeSet in DBpedia.

```
<http://vmuss06.deri.ie/actions#title=Modena&id=383055>
  [...]
  sioca:creates
    <http://vmuss06.deri.ie/changesets#title=Modena&prop=province&date=2009-10-09
        T04:38:53Z>,
    <http://vmuss06.deri.ie/diffs#title=Modena&id=383055&oldid=380059>;
  a sioca:Action.
<http://vmuss06.deri.ie/changesets#title=Modena&prop=province&date=2009-10-09T04
    :38:53Z>
  sioc:has_creator <http://en.wikipedia/User:Plasticspork>;
  cs:changeReason "Change in Wikipedia";
  cs:createdDate "2009-10-09T04:38:53Z";
  cs:subjectOfChange <http://dbpedia.org/resource/Modena>;
  cs:addition _:bnode1;
  cs:removal _:bnode2;
  rdfs:seeAlso <http://vmuss06.deri.ie/DBpediaStats#title=Modena&prop=province&date
      =2009-10-09T04:38:53Z#edits>
  rdfs:seeAlso <http://vmuss06.deri.ie/DBpediaStats#title=Modena&prop=province&date
      =2009-10-09T04:38:53Z#users>
  a cs:ChangeSet.
_:bnode1
  rdf:subject <http://dbpedia.org/resource/Modena>;
  rdf:predicate <http://dbpedia.org/ontology/province>;
  rdf:object "Province_of_Modena";
  a rdf:Statement.
_:bnode2
```

[31] http://purl.org/vocab/changeset (accessed January 2014)

```
rdf:subject <http://dbpedia.org/resource/Modena>;
rdf:predicate <http://dbpedia.org/ontology/province>;
rdf:object "Modena";
a rdf:Statement.
```

Listing 4.7: A ChangeSet for the DBpedia resource "Modena" expressed in Turtle. The object of the property "province" has changed from "Modena" to "Province_of_Modena".

Please note that, in Listing 4.7, the `ChangeSet` instance links with `seeAlso` properties to two resources providing statistical information in RDF about the `dbpedia:province` property. The first one is about the number of edits to this property, on this page, at the time of this ChangeSet. The second one is similar but with the difference that it is about the number of users who edited the property. These statistics are modelled using the SCOVO vocabulary and the resources in this example are explained later in this section in Listing 4.8.

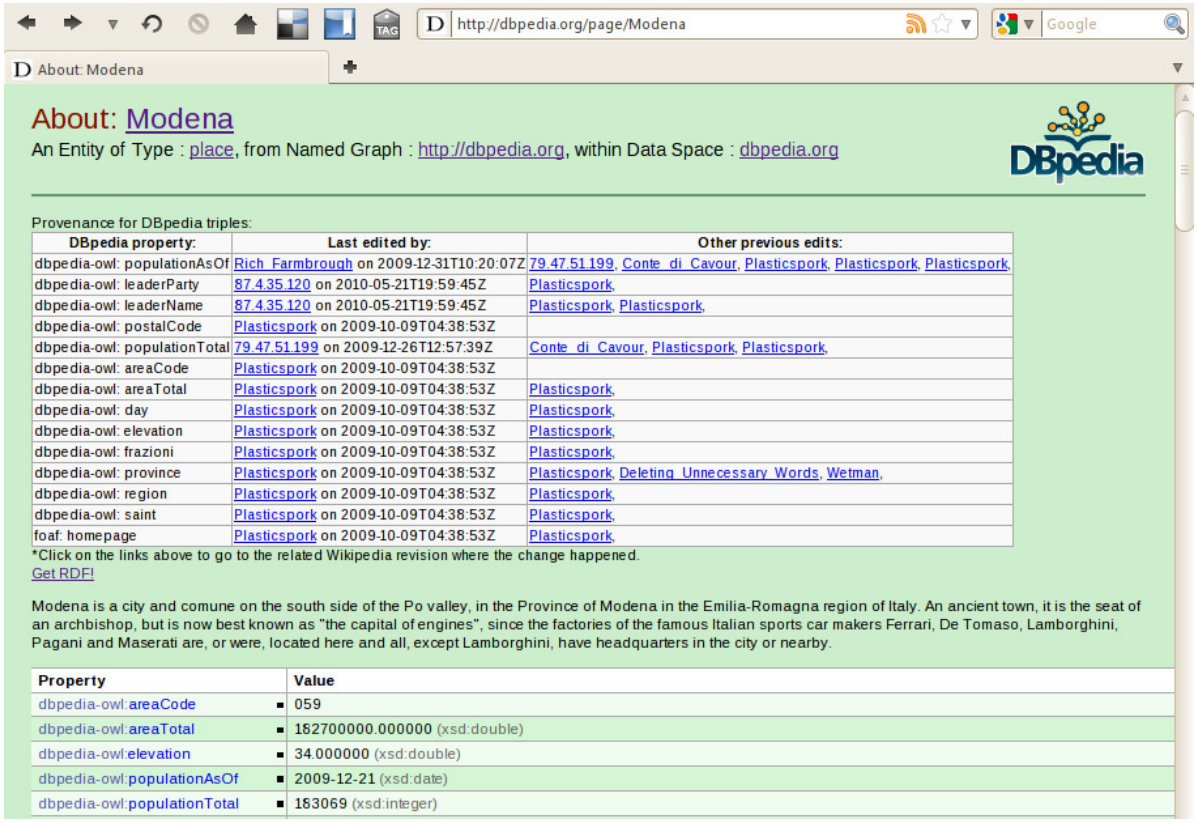

Figure 4.9.: A screenshot of our application displaying provenance information directly on the DBpedia page about "Modena"

Once all the diffs have been analysed, and the related data loaded into the triplestore, we focused our attention on the final part of the application. It is composed by a Mozilla

Greasemonkey script which loads a table on the top of the DBpedia pages based on the results retrieved by another PHP script. The structure of this part of the application is similar to the structure described in Section 4.2.1.4 for the Greasemonkey script running on Wikipedia pages. Similarly, the PHP script receives a request from the Greasemonkey script for a specific DBpedia resource, then it queries the triplestore and replies to the Greasemonkey script with the results embedded in a HTML table. A screenshot of the table displayed on a DBpedia page is shown in Figure 4.9.

In accordance to what we did for the Wikipedia provenance data (Section 4.2.1.4), to make this information about DBpedia also available to machines, we provide these statistics in RDF. Using the SCOVO vocabulary we are able to model, for each property on each DBpedia page, the total number of edits and the number of users contributing to them. In the `scv:Item`s implemented in this case the three `dimension`s involved are:

- the type of information that we want to represent (number of edits or number of users);
- the URI of the DBpedia resource impacted;
- the URI of the DBpedia property involved.

Hence, we created two instances of `scv:Dimension` to represent the first `dimension` (as in the first part of Listing 4.8). The other two dimensions are URIs linked with the `scv:dimension` property (second part of Listing 4.8).

```
@prefix dbst:<http://vmuss06.deri.ie/DBpediaStats#>
@prefix scv:<http://purl.org/NET/scovo#>

dbst:DBPropertyEdits rdfs:subClassOf scv:Dimension ;
  dc:title "Number of edits for the property" .

dbst:Edits a :DBPropertyEdits ;
  dc:title "Number of edits" .
dbst:Users a :DBPropertyEdits ;
  dc:title "Number of users editing the property" .

----
dbst:title=Modena&prop=province&date=2009-10-09T04:38:53Z#edits a scovo:Item ;
   rdf:value 4 ;
   scv:dimension dbst:Edits ;
   scv:dimension <http://dbpedia.org/resource/Modena>;
   scv:dimension <http://dbpedia.org/ontology/province>.

dbst:title=Modena&prop=province&date=2009-10-09T04:38:53Z#users a scovo:Item ;
   rdf:value 3 ;
   scv:dimension dbst:Users ;
   scv:dimension <http://dbpedia.org/resource/Modena>;
```

```
    scv:dimension <http://dbpedia.org/ontology/province>.
```

Listing 4.8: Representing the number of edits and editors of the DBpedia properties with SCOVO

To give a clearer picture of the amount of data generated for this test, we now provide some technical details about the experiment conducted. The total amount of RDF triples generated and stored in our RDF-store is around 770.000. This includes all the provenance data about three Wikipedia articles ("Modena", "Ferrara" and "Ravenna") and other data about the structure of the two categories *"World Heritage Sites in Italy"* and *"Cities and towns in Emilia-Romagna"* and their members. The total number of members belonging to these two categories and all the subcategories is 2645 articles, but for these we did not collect all the revisions, we did that only for the intersection of the two categories. As regards the number of revisions of the three articles collected, each of them has almost 500 revisions.

In terms of time spent for the data acquisition process on a basic single core machine, the total process took around five hours:

- around three hours to get the data from DBpedia and the SIOC-MediaWiki exporter (the slowest part of the acquisition process because of the high number of requests to the Wikipedia API);
- two hours to get all the diffs between the revisions;
- a few minutes to analyse the diffs and match the DBpedia properties.

To better estimate the amount of RDF triples that can be generated by this process, we now provide a comparison between the DBpedia dataset and the result of our provenance extraction process applied to the whole English Wikipedia. In October 2010 the English Wikipedia hosted around 3.5 million articles, with an average number of revisions per article equal to 73.5[32]. Therefore, we approximately consider a total of 257.25 million revisions. Since, we generated with our experiment around 50.98 statements per revision, then for the whole English Wikipedia corpus we would generate almost 13.115 billion RDF triples. Considering that one part of all these statements describes the content and structure of the revisions, and the other part aims at describing their provenance, we estimate that the whole Wikipedia provenance dataset would consist of

[32]From Wikipedia statistics hosted by the Wikimedia Foundation (October 2010): http://stats.wikimedia.org/EN/TablesWikipediaEN.htm

approximately 7 billion triples. As a comparison, the DBpedia dataset[33] consists of 672 million RDF triples out of which 286 million were extracted from the English edition of Wikipedia and 386 million were extracted from other language editions and links to external datasets.

4.4. Provenance for Profiling User Interests

Users on the Social Web interact with each other, create/share content and express their interests on different social websites with many user accounts and different purposes. On each of these systems personal information, consisting of a portion of the complete profile of the user, is recorded. With respect to "complete user profile" we intend the full set of personal information belonging to a person obtained by aggregating the distributed partial user profiles on each Social Web system. Each partial user profile might contain the user's personal and contact information, her interests, activities and social network of contacts. In this work in particular we focus on user profiles of interests as weighted and ranked collections of concepts relevant to the users. All the distributed user profiles on the Web represent different *facets* of the user therefore their aggregation provides a more comprehensive picture of a person's profile [Abel et al., 2010a]. Aggregation of user profiles brings several advantages: it allows for information reuse across different systems, it solves the well-known "cold start" problem in personalisation/recommendation systems (Section 2.4), and provides more complete information to each individual Social Web service. However, the aggregation process is a non-trivial problem which derives from the most popular data integration issues: entity matching and duplicates resolution, conflicts resolution, heterogeneity of the data models of the sources and the consequent need of a common target data model are the most important ones.

Using standard semantic technologies to represent the data sources helps in solving these issues and it provides a unified representation of the target data model. Furthermore a complete semantic representation and management of the provenance of user data addresses the duplicate/conflict resolution issues, since it would allow to track the origins of the data at any point of the integration process [Hartig and Zhao, 2010]. Several approaches for aggregating and representing multi-domain user models have been presented in the state of the art (see Section 2.3) but in most of the cases they are not

[33] DBpedia dataset version 3.6, officially released in January 2011: http://blog.dbpedia.org/2011/01/17/dbpedia-36-released/

aimed at defining a standard, source-independent, architecture that allows for interoperability and integration of profiles of interest on the Web of Data.

Our research on the use of semantics for interlinking social websites and subsequently on provenance on the Web of Data provides us the necessary baseline for our work. In particular we focus on building comprehensive user profiles based on quantitative and qualitative measures about user activities across different social websites. Provenance of data is particularly useful to evaluate on each different website and/or dataset the type and amount of contributions to be attributed to a particular user. For example, this would allow us to infer expertise, interests and qualitative estimations on users' activities.

As argued in [Barbier et al., 2013], it is important to identify provenance attributes on social media that could be vital to the task of identifying provenance of information. Provenance attributes of a user may include name, location, gender, occupation, information content, preferences, etc. These attributes help to "understand" Social Web content, narrow down the possible sources and give more credibility to a piece of information. For example, Barbier et al. in [Barbier and Liu, 2011], show how many attributes of a user can be collected from Twitter alone. In this dissertation we focus on attributes on social media identifying a potential interest of a user.

We investigate Social Web actions (such as comments, status updates, likes, etc.) extracted from popular vocabularies on the Web (SIOC, ActivityStreams, etc. as described in Section 3.3.1) and mapped to popular social media sites. Moreover, we describe how we selected only the ones that help in identifying user interests. We record provenance of the Social Web actions and analyse the impact of the different types of actions on the quality of the user profiles. Our methodology is platform-independent and can be applied to every Social Web system since it is based on the analysis of common Social Web actions, *i.e.* through the analysis of messages or other social networking activities such as comments, places checked-in, liked links, etc. The resulting user profiles consist of entities and concepts potentially representing interests, activities and contexts of the users based on their content generated on the social networks. We use DBpedia resources to represent user interests and determine a score to measure their prominence based on particular heuristics (more details in the next Chapter 5).

In order to determine interests of users, we studied the different types of actions that can be performed on popular social media sites such as Twitter, Facebook and Wikipedia. We selected these social media sites for our investigation as they are among

the most popular ones on the Web, therefore it is easier to evaluate our experiments through user studies. In addition, each one of them represents a different type of social media service. For our analysis we explored popular online vocabularies describing Social Web activities and online user interactions. In particular we analysed the vocabularies offered by the SIOC project and Activity Streams (see Section 3.3.1). Both projects published vocabularies for describing Social Web actions and content, hence our goal was to identify the actions and content features that are applicable to our particular use case. Therefore, from this subset of features, we needed to identify only the ones that are suitable for mining possible entities of interest. In Table 4.3 we summarize the actions and the features that we can use for profiling user interests on the three types of social media.

	on Facebook	**on Twitter**	**on Wikipedia**
Implicit Interests	- comments - status updates - direct post to friend - checkins - media object actions (e.g. post of a video)	- user posts - user replies - retweets - followees' posts - favourite tweets - lists	- text edit - infobox/link edit - "Talk" page edit - article creation
Explicit Interests	- profile: education - profile: workplace - profile: interests - likes		- article creation - add to "watchlist"

Table 4.3.: Social Web actions and content features for mining user interests. These features can indicate an interest "explicitly" or "implicitly".

The features listed in Table 4.3 can be directly retrieved or extracted from the structured or textual content retrieved from the Facebook, MediaWiki and Twitter APIs. Most of these features are all connected to textual information that we can analyse using natural language processing tools in order to spot resources of interest. As shown in Table 4.3, some features *explicitly* express an interest on an entity and some others *implicitly*. The implicit interests carry a higher degree of uncertainty about the spotted entity. This can be either because of the way the entities of interest are spotted (*e.g.* for comments and posts we need to use NLP tools), or because the social action involved does not necessarily imply an interest (*e.g.* Twitter lists or Wikipedia edits).

Additionally, we can consider also another interesting factor to differentiate the types of interests: currently more and more services allow other users to provide information about our own interests. For example, on LinkedIn users can identify others as experts on particular topics (action called "endorsement"). This could be then considered as explicit information about our interests which is however provided by others. We consider this factor as part of our future investigation as these types of actions are currently growing.

One of our goals is to analyse the impact of provenance information and different types of Social Web features on automatically generated user profiles of interests. We will provide a description of the complete profiling process in the following chapters.

4.5. Conclusions

In this chapter we demonstrated, through real examples and practical experiments, how provenance of data plays a crucial role in social media and the Web of Data and especially for user profiling. We described how provenance can be recorded and represented on the Social Web, and consequently used on the Linked Data cloud to track the origins of particular statements and data records. At the same time, provenance on the Web of Data can be used in many different use cases supporting Social Web users and applications. Therefore, provenance of data is the fundamental connection between the Social Web and the Web of Data and it fuels with useful information every step of our profiling methodology (Figure 1.2). It is collected directly from social media together with the user data, then it is represented with our modelling solution using popular ontologies. The potential of semantic representation of provenance in profiling user interests will be shown in the next chapters. One of our goals is to analyse the impact of provenance information and different types of Social Web features on automatically generated user profiles of interests. In this regard, at the end of Chapter 4, we identified potential provenance features that could be used for enhancing our user profiling methodology. In particular, in the next Chapter 5 we describe how we model the described provenance information extracted from different social media platforms (we use the popular Semantic Web vocabularies such as SIOC, as showed in this chapter for the Wikipedia use case). Next, we will detail how we integrate this provenance information in our user profiling methodology.

Chapter 5

Mining User Interests on Social Web Data

5.1. Introduction

User profiling techniques have mostly focused on retrieving and representing user knowledge, context and interests in order to provide recommendations, personalise search, and build user-adaptive systems. However, building a user profile on a single social network limits the quality and completeness of the profile, especially when interoperability of the profile is key and its reuse on different sites is necessary for providing other types of personalisation. Indeed recent studies have shown that users on the Social Web often use different social networking sites for diverse, and sometimes non-overlapping, purposes and interests. For example, websites such as Twitter and Facebook are widely used as news delivery systems or for keeping in contact with friends respectively. In this chapter, we describe the core of our methodology for the automatic creation and aggregation of interoperable and multi-domain user profiles of interests. In particular, we detail how we aggregate data that can be extracted from social media (Chapter 3) along with its related provenance information (Chapter 4). We use a particular semantic modelling solution for the aggregated data that facilitates the integration, manipulation and analysis of the data (see Section 5.2). After that, we introduce several dimensions and heuristics that can be used for mining user interests on top of the aggregated data (Section 5.3). In order to evaluate the heuristics, we propose a user study on different user profiling techniques for social networking websites in general, and for Twitter and

Facebook in particular (Section 5.4). In this regard, based on the results of our user evaluation, we investigate:

1. the accuracy of different methodologies for profiling user interests,
2. the effect of different provenance-based dimensions and heuristics on mining and ranking user interests,
3. the benefits of merging different user models using semantic technologies, and
4. the need for privacy on the Social Web and the design of a privacy-aware management system for user profiles.

We chose to evaluate our profiling techniques on two different social media platforms in particular, Facebook and Twitter, as they are currently among the most popular ones on the Web and therefore would make it easier to find participants for our user study. However, the same techniques and considerations can be applied and extended to other social websites. We implemented our methodology for running experiments and developed a platform-independent architecture which can be applied to every Social Web system. This is mainly because it is based on the analysis of metadata and text produced by the user, *i.e.* through the analysis of messages or other social networking activities such as comments, places checked in, liked links, etc. In particular for our experiments we implemented a system that computes profiles harvested from Twitter and Facebook user accounts. The resulting user profiles consist of entities and concepts representing interests, activities and contexts of the users. We use DBpedia resources and categories to represent the entities included in a profile and we rank the relevance of the interests according to weights computed by different algorithms described and evaluated in the next sections (Figure 5.1). The evaluation of the system and the different methodologies has been conducted with two user studies, the first with 21 participants and the second one — more focused on provenance features for profiling — with other 27 participants.

We conclude the chapter with a brief description of a management system for privacy preferences on user profile data (Section 5.5). Since in this thesis we propose solutions for mining and aggregating personal data from the Social Web, it is necessary to highlight the importance of providing users with tools that would help in protecting and managing their own data. Through a user study we show the need of users for this kind of tools. We propose a system that allows users to define fine-grained privacy preferences over their automatically generated user profiles. The system has a distributed architecture and

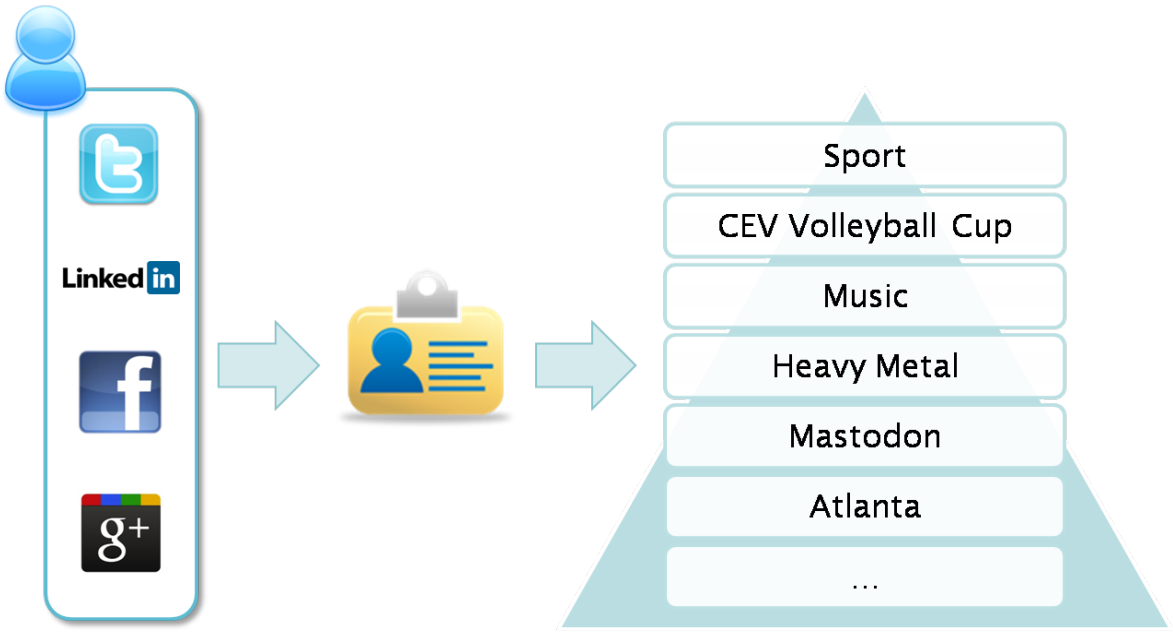

Figure 5.1.: Generation of a user profile of entities, ranked by relevance, extracted from multiple social media sources

it is based on standard Semantic Web technologies, hence it maximises interoperability and flexibility of adoption.

5.2. Extraction and Representation of User Models

During the past few years we have experienced a consistent increase in popularity for Web applications using or collecting data on their users and their behaviour in order to provide adapted and personalised contents and services. This caused the need for exchange, reuse, and integration of their data and user models. As described in Section 2.3, recent relevant studies of the state of the art for this field are compared in [Carmagnola et al., 2011] and in [Torre, 2009], where the authors focus on adaptive systems adopting Semantic Web technologies.

Interesting research that combines user information retrieval/profiling and the Semantic Web has been presented by Szomszor et al. [Szomszor et al., 2008]. The authors investigate the idea of merging users' distributed tag clouds to build richer profile ontologies of interests, using the FOAF vocabulary and matching concepts to Wikipedia categories. In [Carmagnola, 2009] Carmagnola et al. describe one of the most advanced

user modelling systems adopting semantic technologies. The use of RDF for representing user models, and the reasoning capabilities implemented on top of the user models in order to obtain automatic mapping between heterogeneous concepts, are the strongest points of their implementation. Moreover, an extensive approach for ontology-based representation of user models was proposed by Heckmann et al. by introducing GUMO [Heckmann et al., 2005b], a General User Modeling Ontology for the uniform interpretation of distributed user models.

As regards user profiling on social networks, the work presented in [Tao et al., 2011] and [Abel et al., 2011a] shows an interesting and similar approach for creating RDF-based user profiles on Twitter according to the frequency of the entities extracted from the user's tweets. The profiles are then modelled primarily using the FOAF vocabulary. Particularly relevant is the fact that the authors demonstrate the benefits of the amalgamation of multiple Web 2.0 user-tagging histories in building personal semantically-enriched profiles of interest. An analysis of different temporal patterns and dynamics for Twitter user profiles is also provided by the same authors in [Abel et al., 2011b]. Relevant and similar work by the same authors in [Abel et al., 2011d] and [Abel, 2011] is focusing on aggregation of user profiles in general. Hence, they propose an approach for merging different Social Web profile attributes such as workplace, email, phone number, homepage, profile picture, etc. Also, tag based user profiles of preferences aggregated from different Social Web services are evaluated in a tagging recommender system. However, as a comparison, in our work we focus more on investigating different profiling methods and provenance-based heuristics for identifying user interests. We propose entity-based user profiles which can be semantically enriched and connected to the Web of Data and evaluate them through different user studies.

Other related work has been published in [Stan et al., 2011] where the authors describe a system for people recommendation based on *User Interaction Profiles* built extracting entities and keywords from user posts on social networks (from Twitter, in their experiment). A similar architecture for the generation of the profiles is proposed and disambiguation and concept expansion is also done using DBpedia and semantic technologies. On the other hand, an evaluation of the system and the profiling algorithm is not provided and temporal features of user posts are not considered. Despite the interesting combination of traditional content analysis techniques with semantic technologies, in this work the focus is more on building a framework for people recommendation during Web navigation.

Interesting research on Semantic Web applied to user modelling and personalisation has been done by Aroyo et al. [Aroyo and Houben, 2010]. In this work the authors highlight the challenges they see in the near future for user modelling and the adaptive Semantic Web. Furthermore, a review of the research in this field is provided. It is important to note that some of the systems for user model interoperability implement their reasoning capabilities on top of the user data not using Semantic Web technologies but using non-standard application-specific algorithms, making interoperability with other systems more difficult to achieve. For more details on the state of the art for semantic technologies and user modelling we refer to Section 2.3.4.

The steps involved in our profiling methodology for the extraction and generation of user profiles from social networking websites can be summarised with the following main stages. *First*, the data extraction from each specific social networking service and the subsequent generation of application-dependent user profiles. After this phase the *next* steps involve the representation of the user models using popular ontologies, and then, *finally*, the aggregation of the distributed profiles. In this section we describe our RDF modelling solution for multi-domain user profiles of interests and we detail how we integrate user data with the Web of Data and in particular DBpedia. Semantic Web technologies and standard ontologies are the main supports for the development of interoperable services, and these standards make it easier to connect distributed user profiles.

5.2.1. Representing User Profiles of Interest

Our solution for modelling profile data is mainly based on the SIOC and FOAF vocabularies, as described in Chapter 3. Especially FOAF, being one of the most popular lightweight ontologies on the Semantic Web, is used as a basis for representing users' personal information and social relations. Hence, it eases the integration of heterogeneous distributed user profiles. As detailed in Section 3.3.1.1, a FOAF profile consists of a FOAF `PersonalProfileDocument` that describes a `foaf:Person`: a physical person that has several properties describing her and holds online accounts on the Web. However, especially in our case, an important part of a user profile is represented by user interests. In this work we focus in particular on this part of a profile: on how to automatically retrieve interests from social networking sites and how to compute weights expressing their relevance. In Listing 5.1 and Figure 5.2, we display an example of an interest (a `WeightedInterest`) about the entity "*Semantic Web*" with a weight of *0.5*

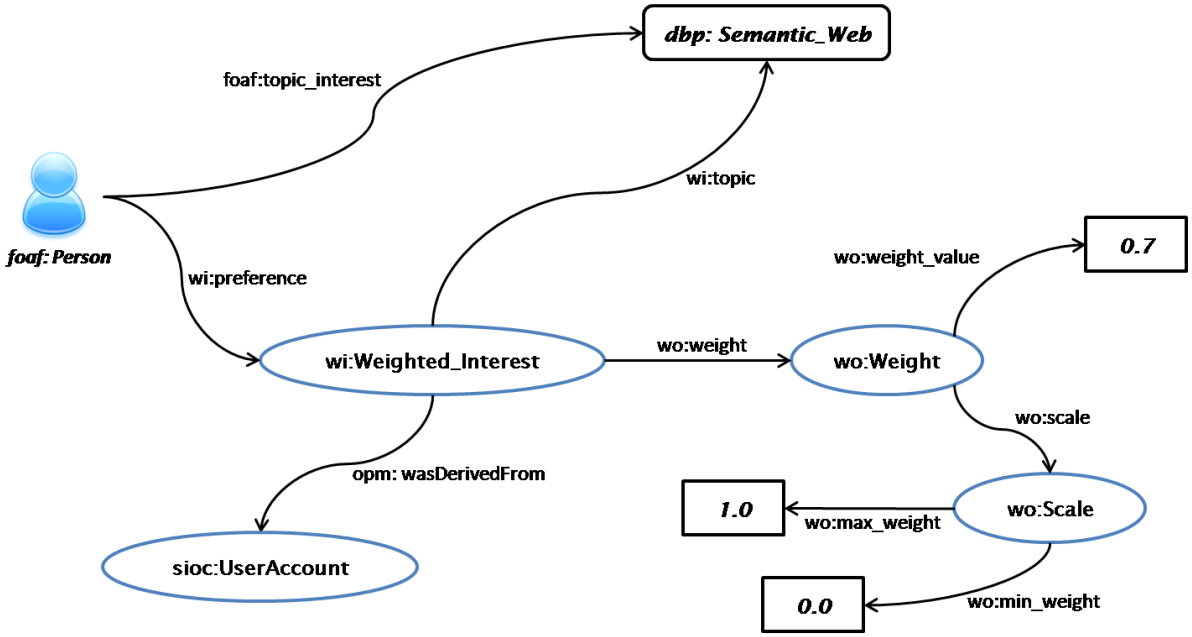

Figure 5.2.: Example of our modelling solution for user interests

on a specific scale (from 0 to 1) using the *Weighted Interests Vocabulary* (WI)[1] and the *Weighting Ontology* (WO)[2]. In order to compute the weights for the interests common approaches are based on the number of occurrences of the entities, their frequency, and possibly some additional factors. These factors might depend on whether or not the interest was implicitly mined or explicitly showed by the user, or depending on a time-based function which computes the decay of the interests over time, or based on the trustworthiness of the social platform, etc. In Section 5.3 we describe the different weighting schemes and heuristics adopted and experimented in our work.

```
@prefix ##please visit http://prefix.cc for the prefixes##

<http://example.org/fabrizio#me>
   a foaf:Person ;
   foaf:name "Fabrizio Orlandi" ;
   foaf:topic_interest <http://dbpedia.org/resource/Semantic_Web> ;
   wi:preference [
      a wi:WeightedInterest ;
      wi:topic <http://dbpedia.org/resource/Semantic_Web> ;
      rdfs:label "Semantic Web" ;
      wo:weight [
         a wo:Weight ;
         wo:weight_value 0.5 ;
         wo:scale ex:01Scale
```

[1]WI Specification: http://purl.org/ontology/wi/core (accessed January 2014)
[2]WO Specification: http://purl.org/ontology/wo/core (accessed January 2014)

```
        ] ;
      wi:appear_time [
         a time:Instant ;
         time:inXSDDateTime "2013-10-16T11:30:00+01:00"^^xsd:dateTime
      ]
      wi:appear_time [
         a time:Instant ;
         time:inXSDDateTime "2013-11-05T02:18:00+01:00"^^xsd:dateTime
      ]
      opm:wasDerivedFrom <http://twitter.com/BadmotorF> ;
      opm:wasDerivedFrom <http://www.facebook.com/fabriziorlandi> ;
   ] .

ex:01Scale a wo:Scale ;
   wo:min_weight 0.0 ;
   wo:max_weight 1.0 ;
   wo:step_size 0.1 .

<http://twitter.com/BadmotorF> a sioc:UserAccount .
<http://www.facebook.com/fabriziorlandi> a sioc:UserAccount .
```

Listing 5.1: A RDF/Turtle representation of an interest (Semantic Web) and its weight (0.5) extracted from two sources at two different time instants

Similarly to the modelling solution described in the previous Chapters 3 and 4 for wikis and their provenance information, we adopt a model based on FOAF and SIOC with the addition of other specific vocabularies for the detailed description of user interests. This solution provides high integration capabilities between all our modelling solutions and allows for trivial interchange of information across heterogeneous social platforms. In the example of Listing 5.1 we use the property `appear_time` and an `Instant` class from the W3C Time Ontology[3] to describe in a generic way a particular time instant when the interest originated from the user's social activity. More detailed temporal dynamics (such as time intervals) and context (such as events and places) related to the interests can be expressed using the `InterestDynamics` class and related properties, as detailed in the WI Ontology specifications[4].

5.2.2. Leveraging Provenance of User Data

Provenance of data is important in this context as it allows data consumers to understand the origins of the interests (time- and source-wise) which are the result of an integration

[3] Time Ontology in OWL. W3C Working Draft 27 September 2006: http://www.w3.org/TR/2006/WD-owl-time-20060927/ (accessed January 2014)

[4] We refer to the WI Ontology specifications for more details about modelling Social Web user interests: http://smiy.sourceforge.net/wi/spec/weightedinterests.html (accessed January 2014)

process. Some data consumers might want to give more relevance to some data sources rather than others according to particular trust measures or differences in contexts and use cases. Moreover it would be possible to recompute new aggregated weight values based on different weighting-schemes and the original data, or enforce privacy rules on the user data based on particular preferences. As regards provenance of the interests, as showed in Listing 5.1, we use the property `wasDerivedFrom` from the Open Provenance Model (OPM) (Section 4.2.1.3) to state that the interest was originated by a specific user account on a website. This property is equivalent to the PROV `wasDerivedFrom` property (see Section 4.2.1.3). In the example in Listing 5.1 we can observe that the interest is derived from both Twitter and Facebook user accounts. As regards to the complete provenance representation of the interests, in the same way as with the wikis in the previous chapter, we record complete information about the origin of an interest. Particularly, in line with the W7 model described in Chapter 4, we connect each *interest* to:

- **(Who)** The agent holding it (a `foaf:Person`);
- **(When)** It's time of creation and modification (with the `wi:appear_time` property);
- **(What)** Its dereferenceable description (using `foaf:topic_interest` and `wi:topic`, and pointing to DBpedia resources);
- **(Which)** The website or user account where it was extracted from (thanks to `opm:wasDerivedFrom`);
- **(How)** The Social Web action which expressed the interest; In this case we can use the same modelling solution proposed in Chapter 4 for the wikis where we use `sioca:Action` (alternatively `opm:Process` or `prov:Activity`) that can be connected to the `WeightedInterest` through the property `sioca:creates` (see Section 4.2.1.2);
- **(Where)** We cannot have related information about a physical location in this case;
- **(Why)** We cannot have precise information about the reason behind an interest.

5.2.3. Interests on the Web of Data

An important aspect of our profiling methodology is the use of entities on the Web of Data to represent the interests of the users. In perticular in this thesis we adopt DBpedia, the semantic representation of Wikipedia. Thanks to its large dataset (around 1 billion RDF triples) and its cross-domain nature DBpedia has become one of the most important and interlinked datasets on the Web of Data [Cyganiak and Jentzsch, 2010]. In this thesis we adopt DBpedia for our use cases and experiments, especially because of its very large and domain-independent knowledge base. However, we could use any dataset offered on the Web of Data for representing interests, even domain specific ones. Our methodology would not change and it is also supporting the adoption of different distributed knowledge bases, even the entire Web of Data. Representing interests using DBpedia resources has two main advantages: integrates the user profiles with the Linked Data cloud, and provides a larger and "fresher" set of terms as compared to traditional taxonomies or lexical databases such as WordNet[5]. In [Ponzetto and Strube, 2007] the authors demonstrate the benefits of using Wikipedia (or DBpedia) for computing semantic relatedness and for named entity representation as compared to WordNet and other knowledge bases. In our work we use DBpedia not only to link to its entities but also to extract related categories for concept expansion and to analyse the structure of the categories graph in order to understand the relevance of a category for representing a user interest. Our plan is to extend this analysis also to other Linked Data datasets and resources.

5.3. Heuristics for Interests Mining on the Social Web

5.3.1. Bag-of-Words vs. Disambiguated Entities

Most of the state of the art methods for user profiling which need to identify possible entities of interest in textual user-generated content employ *tag-based* user profiles [Michlmayr et al., 2007, Abel et al., 2010b]. In other words, the Social Web textual content is analysed and processed with traditional text-processing techniques such as stemming and stop-words removal to identify words or tags that frequently occur in the corpus. User profiles are then sets of frequent tags ranked by tag frequency. This methodology leads to errors as it is not considering the position of the words in the

[5] "About WordNet" 2010. http://wordnet.princeton.edu

sentences, the language grammar, the context of the sentences and possible ambiguities. For this reason we implement and evaluate *entity-based* user profiles in our work and compare them with the tag-based ones. In order to identify entities within the text we use specific tools that offer natural language processing capabilities and named entity extractors that spot entities such as places, persons, organisations, etc. and provide the related DBpedia resources. As described later in Section 5.4.1.2, these tools perform entity disambiguation, as entities are linked to URIs on the Linked Data cloud and ambiguities are resolved analysing the context of the sentences.

5.3.2. Time Decay

Interests change in their relevance for a user over time and in most of the cases preferences that have been expressed by a user only in the past become less relevant than interests which have been expressed very recently. We can state in general that the relevance of interests for a user decays with the time. This condition is verified also in other related studies such as in [Ding and Li, 2005] [Abel et al., 2011b] and [Nakatsuji and Fujiwara, 2012]. With a time decay method we assume that the recent entities of interest extracted from the Social Web activity of a user reflect his/her current interests more than the older ones.

As suggested by [Ding and Li, 2005] and most of the current state of the art, an exponential time decay function is used to compute the relevance of the interests over time. This type of function has been evaluated in the aforementioned related work and it demonstrated its efficiency in many other profiling algorithms. However, [Nakatsuji and Fujiwara, 2012] shows the benefits of adopting different innovative temporal patterns, based on grouping interests into epochs, in order to understand the dynamics of the interests over time. Because of the early stage of this research, and the marginal improvement presented over traditional time decay-based methods, we decided to adopt more traditional time decay approaches for our experiments. This would also facilitate the comparison of our work with other research studies.

We use an exponential decay function to evaluate the relevance of each interest according to its position on the user timeline. The function gives higher weight for interests occurred recently and lower for older interests. Following the aforementioned state of the art studies, we adopt an exponential function as it has been shown to be

simple and effective. The exponential decay function is:

$$x(t) = x_0 e^{-t/\tau} \tag{5.1}$$

Where: $x(t)$ is the quantity at time t, $x_0 = x(0)$ is the initial quantity (at time $t = 0$), $\tau = 1/\lambda$ is a constant called *mean lifetime* and λ is a positive number called the *decay constant*. When an interest reoccurs multiple times we use the average of the timestamps of the different reoccurring events as time t.

Applying this function to our use case, in order to compute the time decay of the interests, we need to arbitrarily choose values for x_0 and τ which are constants of the function. For our experiment we set $x_0 = 1$, the maximum possible value of the function. We also defined an initial time window where the interests are not discounted by the decay function (7 days). Moreover, in order to identify an appropriate value for τ, we decided to choose two possible values and evaluate them with an experiment and a user study. The constant τ represents the time at which the function value is reduced to $1/e = 0.368$ times its initial value x_0. In our experiment we evaluate the following two values: $\tau = 120 days$ and $\tau = 360 days$. From a practical point of view the two values indicate that an interest value is discounted to 37% of its initial value respectively after 120 and 360 days. These two values have been selected following preliminary experiments and the aforementioned related work. In particular, with some early experiments, we identified substantial changes in the rankings of the interests using these two values. Moreover, similar values have been experimented also in [Abel et al., 2011b].

The exponential decay function is directly applied to the frequency value of the interests, calculated as the ratio between the number of the interest occurrences and the total number of occurrences of all the interests. As regards the time considered for the decay function (the value of t), we compute the average time of the timestamps collected for each interest.

An interesting useful distinction that could be implemented is between long-term interests and short-term or occasional interests. Interests reoccurring many times separated by long time periods indicate stable/long-term interests, while the opposite could happen to occasional interests. However, in order to keep the complexity of our experiments low, we did not implement this distinction and we will focus on it in our future work.

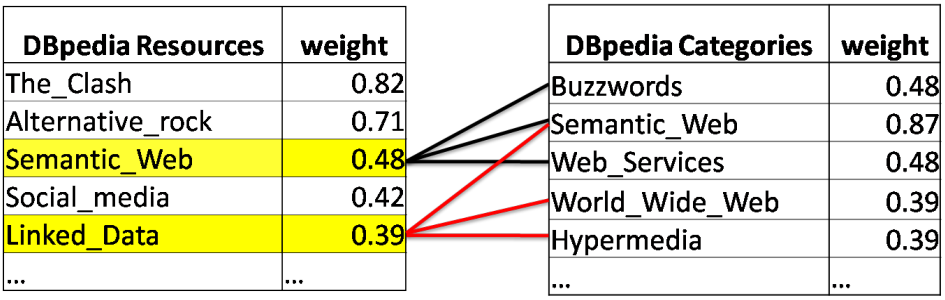

Figure 5.3.: Example of a possible resource-based profile (on the left) with relevance weights and a corresponding portion of a category-based profile (on the right) with recomputed weights.

5.3.3. Categories vs. Resources

In Section 5.2.3 we mentioned that in our methodology we link every entity or concept representing a user interest to the Web of Data, and in particular to DBpedia. In this regard, on DBpedia we can have two main different types of resources: either standard resources — which correspond to entities or pages on Wikipedia — or categories — which represent groups of resources or Wikipedia articles. Therefore, we note that two different types of user profiles can be created: *resource-based* and *category-based*. The category-based methods implemented in our work extract from DBpedia all the related categories of the DBpedia resources that have been computed with the resource-based methods. As soon as we get a DBpedia entity from the entity recognition tool, this takes part of the resource-based profile. Then, for every resource collected we query the DBpedia SPARQL endpoint[6] to retrieve the categories that are connected to the resources. A DBpedia resource is linked to its categories through the Dublin Core[7] `subject` property. From each category, which is defined as a `skos:Concept`, is also possible to navigate the categories graph to obtain more related categories using the `skos:broader` and `skos:narrower` relationships. This option would be useful for use cases where it is necessary to broaden the user profiles, for instance for recommendation systems. Once all the categories are retrieved from DBpedia starting from the original resource-based user profile, we can create the category-based profile and assign different weights to the categories according to different weighting-schemes. This involves then a second aspect which will be evaluated in our experiment (Section 5.4.2.1).

[6] http://dbpedia.org/sparql (accessed January 2014)
[7] http://dublincore.org/documents/dcmi-terms/ (accessed January 2014)

We developed two different weighting-schemes for the categories. The first one is the most straightforward one: it propagates the weights of the resources computed with any resource-based method to the categories. Hence, the weight of each category is the sum of all the weights of the interests/resources belonging to that category. The idea of the second type of weighting-scheme is to reduce the weight of the category (computed in the same way as the first weighting-scheme) if the category is a too "broad" or generic category, hence it is not descriptive enough or useful for a user profile. More in detail, analysing the structure of the categories on DBpedia we noted that generic categories usually contain many resources or have several subcategories. We then implemented a solution to lower the weight of this type of categories. In this case the discount value that multiplies the original weight of the category is computed as follows:

$$CategoryDiscount = \frac{1}{log(|SP|)} \cdot \frac{1}{log(|SC|)} \tag{5.2}$$

where: SP = *Set of Pages belonging to the Category*, SC = *Set of Sub-Categories*.

The number of subcategories and pages is retrieved again using the DBpedia SPARQL endpoint. This method, for example, discounts the value of too generic categories such as *"Living People"*, which are not meaningful and representative of a user interest. At the same time the method keeps the original weight for relevant and particular categories such as *"RDF"*.

5.3.4. Provenance-based Features

As described in Chapter 4 provenance of data can be useful in understanding the origin and the context of Social Web data. We utilize several provenance features in our profiling algorithm to improve the accuracy of interest mining. In particular, we identify several features in the extracted Social Web data for which we can evaluate the impact or influence on the accuracy of user profiles of interests. In order to determine and score interests of users, we studied the different types of actions that can be performed on social media. In Section 4.4 we described how we can use provenance information for user profiling. In particular, in Table 4.3 we summarized the actions and the features that we collect and use in our "provenance-aware" algorithm. We retrieve and semantically represent the features listed in the table for the users' collected social data. We then analyse how prioritising (or giving more relevance to) certain types of actions affects the

accuracy of the profiling algorithm. In turn, we study if entities extracted from some particular actions lead to better or worse interests for a user profile.

Other provenance-based features that we analyse are: the type of social media source (whether it is a microblog or a wiki etc.), the social media site (e.g. Facebook, Twitter, etc.), the time dimension (see Section 5.3.2 about time decay), and whether the entities of interest are extracted implicitly or explicitly (see Table 4.3).

We evaluate this with a user study with 27 volunteers that we describe in Section 5.4.2.2. The outcome of the user study is then used to tune the values of the heuristics of our user profiling module.

5.4. Aggregated User Profiles of Interests on the Social Web

This section provides a description of the architecture proposed for the automated creation and aggregation of interoperable and multi-source user profiles (Section 6.3.2). We also detail the experiment (Section 5.4.2.1) we conducted in order to evaluate our architecture and the different heuristics for ranking user interests introduced in the previous section. A complete analysis of the experiments and a user study is provided in Section 5.4.2.2. The experiments have been conducted using both Facebook and Twitter as social media sources.

5.4.1. Software Architecture

We implemented a Web service (written in PHP) that requires users to log-in with two of their Social Web user accounts and returns a representation of their user profile of interests in RDF. The generated profile is the aggregated result of the analysis of their activity on such services. From an architectural perspective, the profiling framework is composed of three main modules (Figure 5.4):

(1) Service-specific data collector;

(2) Data analyser and profile generator;

(3) Profiles aggregator.

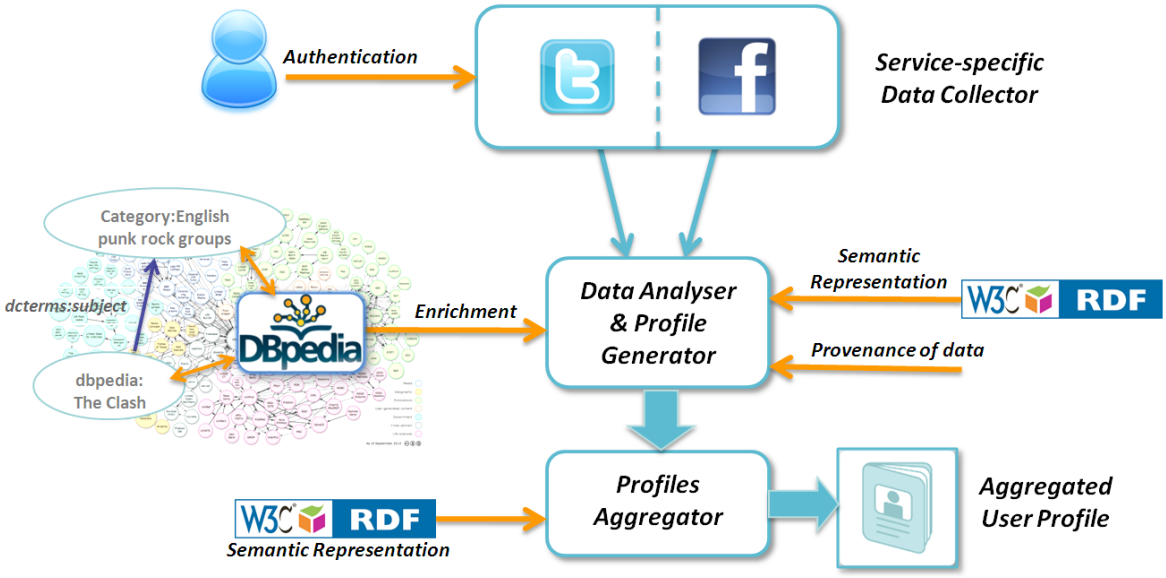

Figure 5.4.: Architecture Diagram

The second and third modules include the representation of the profiles of interests using the modelling solution described in Section 5.2.1. In *module (2)* the semantic representation involves only one — single source — profile, while in *module (3)* RDF is generated for the final aggregated user profile. The implementation of the system for our experiment and evaluation is based on two of the most popular social networking sites: Facebook and Twitter.

5.4.1.1. Service-specific Data Collector

The first module is the module that interacts directly with the source of the profile, the social networking site. This module is responsible for the interaction with the service API, the user authentication, and the data collection from the API. In order to collect private data about users on social websites it is necessary to have access to the data granted by the users. Then it is necessary to request access to the profile data in order to fetch most of the data which is often private by default. In particular, dealing with Facebook and Twitter, we implemented the OAuth 2.0 [8] authentication system required by these platforms to access users' private data. We implemented two distinct modules, one for each social service, each of them including the OAuth authentication system. We

[8] http://oauth.net/2/ (accessed January 2014)

adopted two different libraries for PHP: *Twitter-async*[9] and *Facebook PHP-SDK*[10]. Using the Twitter API we are able to request up to 3,200 of a user's most recent statuses, while Facebook adopts rate limits. The type of data we collected from Facebook is: status messages posted on the user's wall, the entities liked, the places checked-in and user profile information. In the same way on Twitter we retrieve the status messages posted by the user on his/her timeline and other users' messages that the user "retweeted" or favourited. For both Facebook and Twitter we limit the collected information to one year of user history, unless the limit imposed by the social platform occurs.

5.4.1.2. Data Analyser and Profile Generator

Once the user data has been collected from the different platforms the next step is the analysis of the data in order to identify entities and generate the profiles. In this work we use a named entity recognition software to extract entities from the text retrieved at the previous stage. In particular we use Zemanta[11], a Web service that exposes an API and provides text analysis tools to developers. The service in particular offers natural language processing capabilities and a named entity extractor that spots entities such as places, persons, organisations, etc. and provides the related DBpedia resources. It performs entity disambiguation, as entities are linked to URIs on the Linked Data cloud and ambiguities are resolved analysing the context of the sentences[12]. We chose Zemanta for its automated DBpedia URIs suggestion capabilities and for its satisfying performances in analysing short messages such as tweets. According to the state of the art Zemanta, in comparison with similar services such as Alchemy API[13], DBpedia Spotlight[14] and Open Calais[15], performs slightly better than the others. Recent research on this topic [Rizzo and Troncy, 2011] is supporting this statement and suggests Alchemy API and DBpedia Spotlight as the main alternatives. According to the study Zemanta has higher precision than the other tools in recognising named entities and disambiguating them with proper URIs (which is the most important feature for our work). This is supported by a substantial agreement between the evaluators during the experiments conducted by Rizzo et al. According also to other studies [Mendes et al., 2011] Zemanta

[9] https://github.com/jmathai/twitter-async (accessed January 2014)
[10] https://github.com/facebook/facebook-php-sdk (accessed January 2014)
[11] http://developer.zemanta.com/ (accessed January 2014)
[12] Zemanta API companion documentation: http://developer.zemanta.com/docs/ (accessed January 2014)
[13] http://www.alchemyapi.com/ (accessed January 2014)
[14] http://dbpedia.org/spotlight (accessed January 2014)
[15] http://www.opencalais.com/ (accessed January 2014)

dominates in precision but has lower recall than DBpedia Spotlight and the WikiMachine [Bryl et al., 2010] that have similar F_1-scores. To note also that other tools such as Alchemy API perform better in categorisation but this feature is not required in our work since we can use the DBpedia taxonomy for this task. Following the results provided by these publications, we decided to use also DBpedia Spotlight in combination with Zemanta. We use Spotlight in the same way as Zemanta, with the only difference that we use it only when entities spotted by Zemanta have an overall low confidence value. This means that Zemanta does not have enough confidence for most of the entities spotted. For an extensive evaluation of these tools we rely on the work published in [Rizzo and Troncy, 2011] and [Mendes et al., 2011] as this is not the focus of our work. To mention that these tools are continuously improving their performance over the years and according to these studies they can reach precision levels approximately around 80% and recall roughly around 30%, depending on the use case and experiment setup.

In our framework in particular we perform entity extraction algorithm on every message and social activity that we collected at the previous stage. For each message we then record the time the action was performed by the user and the set of entities retrieved for that message. A list of entities (DBpedia URIs provided by Zemanta or Spotlight) is then populated during this phase. For every entity we record the number of occurrences and the timestamps for each of them. Hence, not only the latest occurrence is kept into memory, but also the timestamps for all the previous ones. This part is important for computing the weights of the interests.

In this regard we combine the number of occurrences with a time decay function that evaluates the distribution over time of the interests (as described in Section 5.3.2). The exponential decay function is directly applied to the frequency value of the interests, calculated as the ratio between the number of the interest occurrences and the total number of occurrences of all the interests. As regards the time considered for the decay function (the value of t) we compute the average time of the timestamps for each interest. Following the computation of the weights for all the interests, all the values are then normalised in an interval between 0 and 1.

Finally, the set of interests generated after this second phase has to be represented in RDF according to the modelling solution described in Section 5.2.1.

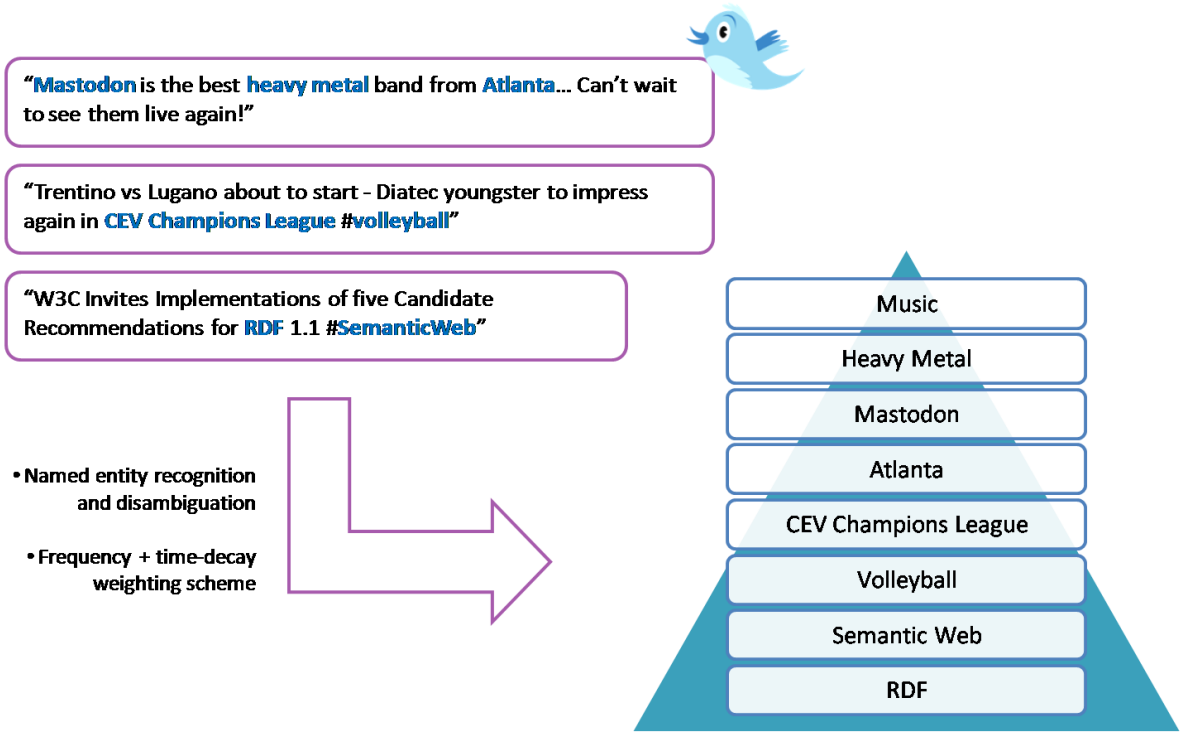

Figure 5.5.: Illustrative example for interest mining from a Twitter feed of messages.

5.4.1.3. Profiles Aggregator

The final phase of the profiling framework is the aggregation of all the single source user profiles. The challenge arising when merging user profiles is the necessity to resolve shared interests reoccurring on different profiles and to recalculate a global weight for these interests. Their new aggregated weight should then be higher than their weight on a single profile, as reoccurring concepts on different social media sites indicate a strong interest. If the same interest is present on two or more profiles it is necessary to: represent the interest only once, compute its new global weight, and update the provenance of the interest keeping track of the sources where the interest was derived from. As regards the computation of the aggregated global weight for the interest generated by multiple sources, we propose a simple generic formula that can be adopted for merging the interest values of many different sources. The formula is as follows:

$$G_i = \sum_s W_s * w_{is} \tag{5.3}$$

Where: G_i = global weight for interest i, W_s = weight associated to the source s, w_{is} = weight for the interest i in source s.
Using this formula it is possible to specify static weights associated to each source depending on which source we want to give more relevance to. In our particular experiment we did not assign different weights to Twitter and Facebook. We considered every social website equally in terms of relevance, hence we multiply each of the two weights by a constant of 1/2 and then we sum the results. The following formula summarises the computation of a new global weight (G) as result of the two original weights (W_1, W_2). It is the same formula that we propose in the previous section (formula 5.3) with the following values: $W_s = 1/2 \; \forall s$. Hence: $G_i = 1/2 * w_{i1} + 1/2 * w_{i2}$.
The different values associated to W_s depend on the particular source but can also be associated to a type of source. For example microblogging platforms (e.g. Twitter, Identica, etc.) could be associated with the same value. To note that this fine-grained weighting strategy is dependent on the particular application for the user profiles or on the users themselves.

5.4.2. Evaluation of Aggregated User Profiles

5.4.2.1. Description of the Experiment

This section describes the experiment that has been conducted in order to evaluate the implementation of the system and different aspects and methodologies of user profiling. The first aim of this experiment is to evaluate the accuracy of aggregated user profiles in relation to the weighting-scheme and the ranking of the interests. The system allows users to generate user profiles from their Twitter and Facebook user accounts. In particular at this stage we generated 6 types of user profiles which differ for the following aspects: (i) The type of DBpedia entities adopted (either Categories or Resources). (ii) The type of weighting-scheme for category-based methods (two different methods). (iii) The type of exponential decay function (either with a shorter time decay parameter $\tau = 120$ days, or a longer one $\tau = 360$ days).

The listed types of profiles have been described in Section 5.3 The third and last aspect chosen for our experiment is the exponential time decay function applied to the computation of the weights. As explained in Section 6.3.2 we chose two values of time decay parameter (120 and 360 days) and implemented all the three different methods two times with the two decays. Hence, in conclusion, for each user we ran our experiment

with 6 different profiling algorithms: *resource*-based profiling, *category*-based profiling *1st* method and *category*-based profiling *2nd* method, each of them twice because of the two time decay parameters (we use the following abbreviations: *Res 360, Res 120, Cat1 360, Cat1 120, Cat2 360, Cat2 120*). The generation of the 6 user profiles takes from 6 to 9 minutes on a standard dual core laptop. In Table 5.1 we display the average number of interests generated for each method. This has been evaluated with 21 users and, on average, using category-based methods generates 6.8 times more interests than the resource-based ones, and the longer time decay (3 times longer) generates 1.4 times more interests.

Res 360	**Res 120**	**Cat 360**	**Cat 120**
44.5	33.1	308.1	221.8

Table 5.1.: Average number of interests, per user, per profiling method

In this section we analyse the evaluation of the implemented system and the different methodologies proposed. In order to evaluate the validity of our approach for generating aggregated user profiles we conducted a user study with 21 users. Demographics include users from 21 to 45 years old, all of them proficient with Social Web systems and 76% of them working/studying in information technology fields. The survey we proposed to the users is composed of 10 questions and the average time taken by the users to complete it was between 9 and 10 minutes. Table 5.2 shows their answers for: "How often do you *actively* use Facebook/Twitter? (i.e. post a message/link, press "like" buttons, check-in, etc.)". From the table is clear that in general our sample uses more actively Facebook than Twitter.

-	**Facebook**	**Twitter**
every day	**66.7%** (14 users)	14.3% (3 users)
every other day	19.0% (4 users)	14.3% (3 users)
once/twice a week	9.5% (2 users)	23.8% (5 users)
once every two weeks	0.0% (0 users)	**28.6%** (6 users)
once a month	4.8% (1 user)	19.0% (4 users)

Table 5.2.: Active usage of Facebook and Twitter

The second type of question we asked users was about enumerating a list of entities and concepts that they were expecting to be representative of their interests, activities

-	Cat1 360	Cat1 120	Cat2 360	Cat2 120	Res 360	Res 120	Baseline
Average Score	5.67	5.20	5.49	5.26	**7.24**	6.81	3.46
No. of Non-Relevant	31	42	34	46	**21**	22	74
Tot No. of Scores	210	209	209	210	**210**	205	200
Precision	0.857	0.799	0.837	0.781	**0.900**	0.893	0.630
MRR	0.921	0.937	1.00	0.933	**1.00**	1.00	0.858
P@10	0.852	0.800	0.838	0.781	**0.900**	0.895	0.610

Table 5.3.: Statistics about the user study for each of the 6 profiling methods and the baseline.

and context on both Twitter and Facebook. This question helps understanding if the topics expected by the users are represented also in the user profiles that we generated. Using the answers to these questions we were able to identify the interests in the generated profiles that were relevant to the users and the interests that were expected by the users but missing in our profiles. This allowed us to compute an approximate *recall* value for our profiles, even though this method might not be very accurate since users had no restrictions in choosing their expected interests. Also, since users do not have perfect memory, we acknowledge the fact that this recall measure is just an estimation and an accurate recall value in this case cannot be computed. The computed average recall value for all the profile types is: 0.740. Next we evaluate the precision according to different measures.

The other remaining 6 questions were all similar, and required users to give a relevance score to each of the top 10 interests for each of the 6 proposed profiling methods. For each method we provided a table of ten interests ordered by weight. The exact question formulated to the users was: "Consider Table X. Please rate how relevant is each concept for representing your personal interests and context.". The options available to users for rating the interests were the following: *0 (not at all or don't know), 1 (low relevance), 2, 3, 4, 5 (high relevance).* Users were then rating the interests on a scale from 0 to 5 (rescaled then in values between 0 and 10) and they were supposed to give a score equal to 0 in case the interest was totally unrelated or unknown.

5.4.2.2. Evaluation and Results

To evaluate the results obtained from the user study, in Table 5.3 we summarise the values obtained for each of the 6 methods considering as a non-relevant result the case

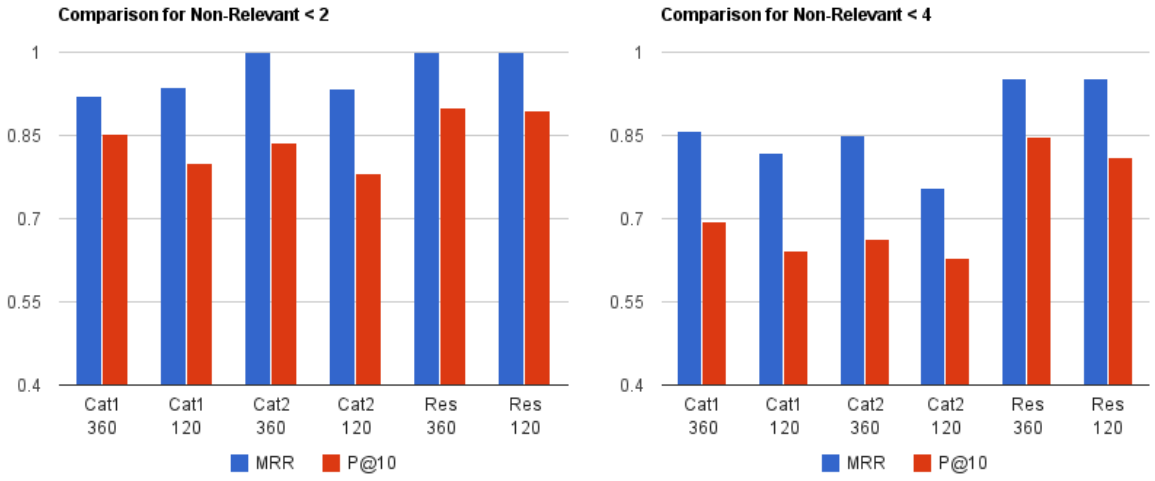

Figure 5.6.: User Evaluation - MRR and P@10

when the rate value is 0 (so the non-relevant value is below 2 in a 0 to 10 scale). The values of the average score are on a 0 - 10 scale. We use the Mean Reciprocal Rank (MRR) and the Precision at $K = 10$ (P@10) statistical measures to evaluate the accuracy of the profiles and the ranking/weighting scheme. The MRR statistic indicates at which rank the first item relevant to the user occurs on average. The P@10 measure indicates the mean probability that a relevant item occurs within the top k of the ranking. In Figure 5.6 we can see a comparison of the MRR and P@10 values calculated both for the case that considers non-relevant an interest with score lower than 2, and for the other which considers non-relevant scores lower than 4.

As regards the statistical significance of the results, we tested our data with both a Wilcoxon's matched pairs test and a paired two-tailed t-Test. The first one is more appropriate because it is a non-parametric method and also our sample is relatively small. Yet, both the tests provide the same results. We tested the differences between the three main methods (and especially the differences between category-based methods and resource-based ones) we calculated p values lower than 0.05, which confirms the significance of the results. As regards the comparison between the samples with two different τ values the Wilcoxon's test rejected the hypothesis of statistical significant difference between the two samples but the computed p values are very close to the α value ($\alpha = 0.05$). This means that we cannot state that the results for those cases are significant, although those numbers are not very high. However, we probably have to increase the number of users in the sample in order to test whether the theory is statistically valid or not.

Bag-of-Words vs. Disambiguated Entities As we can see all the values of MRR and P@10 are satisfying and encouraging. As a comparison with traditional non-semantic approaches in Table 5.3 we also included a "*Baseline*" method. This method is a simple traditional approach, a tag-based user profiling method. It retrieves the most frequent words from the user posts and ranks them according to their number of occurrences. Stemming is applied in our case and stopwords are also removed. As showed in Table 5.3 this method performs clearly worse than all the other entity-based methods. Precision measures for tag-based profiles are roughly around 0.6, meanwhile the other entity-based methods have around 0.8 or 0.9 values of precision. The evaluation has been completed by almost all the users who evaluated also the other methods.

Time Decay Further, the two methods using DBpedia resources, and not the categories, perform better than the others using categories, and at the same time the results for $\tau = 360days$ are slightly better than for $\tau = 120days$. Therefore we would infer that a longer time frame, and a smoother exponential decay function, would better represent users' interests. To note that this is probably true in cases similar to this one, where the aim of the profile is to globally represent user interests and contexts, but it might not be true in cases such as news recommendations where a "fresher" and updated user profile might perform better (see next Chapter6 and the work in [Abel et al., 2011c]).

Categories vs. Resources Interesting to note that DBpedia resources are slightly more precise and specific for building profiles than the related categories extracted, however the results obtained using categories are very close to the traditional methods using just DBpedia resources. Moreover, as an advantage for using categories, as we have shown in Section 5.4.2.1, the number of categories that can be extracted for profiling a user is almost 7 times larger than the number of resources. This is particularly useful in recommendation use cases, where there is a need of getting as much related concept as possible for profiling a user. Further, according to the results, we think that mixed approaches adopting both categories and resources for user profiling can be highly beneficial and need to be investigated. According to users' feedback during the survey, DBpedia resources revealed to be often very specific and narrow, so not always appropriate for representing user interests. On the contrary, the categories for the first method were sometimes too generic (e.g. the frequently occurring *"Category:Living_People"*) and although the second category-based method is capable of removing the very broad

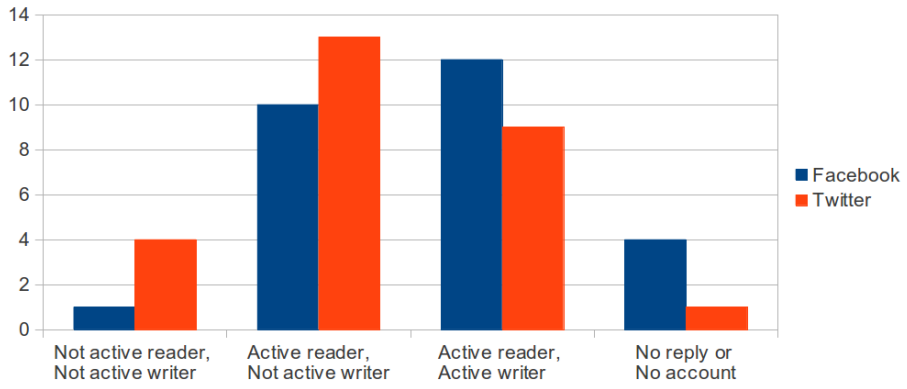

Figure 5.7.: Distribution of the level of activity of the participants on the two social networks for the second user study.

categories from the top of the interests' list, it has the problem of introducing more noise.

Provenance-based Features In order to further evaluate the validity of the user profiling methodology and the provenance-based features, we designed an additional and more extensive user study. Similarly to the previous user study, we again asked users to provide feedback on their own user profiles. This survey has been conducted online with 27 participants: 9 females and 18 males, 4 of them between 18 and 25 years old, 17 between 26 and 33 years old, 4 between 34 and 40 years old and only 1 between 41 and 50 years old. The survey was anonymous and consisted some general questions about their generalities and the average amount of activity on the social networks. The main difference with the previous user study is that we asked people to authenticate to our online prototype, generate their user profile of interests and then rate 30 of their automatically generated entities of interest. The methodology for the generation of the user profiles in this case is only the "*Res 360*" resource-based method with $\tau = 360 days$ for the time decay function (as previously described in this section). We chose this method because it resulted as the best performing one for this type of evaluation. Figure 5.7 displays the distribution of the number of participants for the reply to our question about their average level of activity on Twitter and Facebook.

As for the results of the ratings of the entities of interest we collected 30 marks for each user, so in total 810 ratings (529 distinct entities). In Table 5.4 we summarise the results evaluating the performance of the profiling algorithm. We display the average mark given by the users for both the top 30 and top 10 interests ranked using our

	AVG Score	std.dev.	P@k t>1	P@k t>2	P@k t>3
Top k = 30	3.35	1.47	0.804	0.677	0.525
Top k = 10	3.61	1.49	0.826	0.722	0.622

Table 5.4.: Average user scores (1 to 5 scale) and precision for the profiling algorithm

"*Res 360*" weighting strategy. Users were asked to mark the relevance of each entity of interest with an integer value between 1 and 5 (1 being not relevant and 5 very relevant). The mark is given according to how much each user perceives an entity as a personal interest. We also evaluate the precision of the profiling algorithm, considering the number of relevant interests provided in the top 30 and top 10 lists. To do so, we considered different thresholds in the multipoint scale used to evaluate the interests. In Table 5.4 we show the results obtained for the *precision at k* (*P@k*) where k equals 30 and 10 and $t{>}x$ means that we consider an interest as being relevant if it is marked x *or higher*.

As we can see from this evaluation the results are quite satisfactory especially compared to the results we obtained for a method based on a bag-of-words approach (tag-based user profile, as described in Section 5.3.1). For this method, that we use as a baseline, we obtained a precision at 10 equal to 0.610 for $t > 1$. This is in line with the results obtained with the previous user study (Section 5.4.2.2 and Table 5.3).

To note that even with this experiment it was still not possible to compute an accurate recall value, hence we asked the users in our survey to estimate a coverage percentage for the top 30 interests. In other words, after the users evaluated the 30 interests, we asked them to choose an approximative percentage representing the coverage of those interests compared to their full personal set of interests. The results for this question are: 8 participants declared that the 30 interests covered less than 40% of their total personal interests; 9 users declared between 40% and 60%; 8 users between 60% and 80%; and 2 users more than 80%.

Interesting is the outcome of the study about the impact of provenance of data and the different types of Social Web features considered over the quality of the user profiles. *What are the best social features and sources of user data that we should consider for mining user interests?* To answer this question we recorded provenance information of the collected user data and analysed the average user marks given to the entities that were extracted from the features listed before in Table 4.3. The results of this analysis are

Social Feature	AVG Score	std.dev.
FB education	4.62	0.49
FB workplace	4.60	0.57
TW followees posts	4.03	1.23
FB checkins	3.95	1.25
FB interests	3.95	1.57
FB likes	3.92	1.31
TW favourite posts	3.76	1.28
TW retweets	3.76	1.35
TW posts	3.61	1.34
TW replies	3.52	1.41
FB status updates	3.50	1.53
FB media actions	3.24	1.48
FB comments	2.56	1.54
FB direct posts	2.37	1.59

Table 5.5.: Average user scores associated to each type of Social Web feature

	AVG Score	std.dev.
Explicit Interest	4.27	0.98
Implicit Interest	3.43	1.40

Table 5.6.: Average user scores associated to each group of implicit/explicit features (on a 1 to 5 scale)

summarised in Table 5.5, for every Social Web feature we report the related average user score and standard deviation (we use FB as shortener for Facebook and TW for Twitter). The same procedure for *explicit* or *implicit* features shown in Table 5.6 aggregating the scores of the single features into two groups. As expected explicit interests provide better scores than implicit ones, but implicit ones are necessary for extending the number and range of the entities extracted. We note that features such as *workplace*, *education* history on the profile information and *checkins* have high scores and are all connected to places as entities of interest. Moreover, entities extracted from tweets received by *followees* are more accurate than those from the *posts* of the user itself. The lowest accuracy is obtained by Facebook comments and directed posts on friends' wall, clearly because of their very noisy nature.

The outcome of this study has been directly implemented as an extension or improvement of our profiling algorithm. In particular for the evaluation of our user profiling methodology on recommendation systems (described in the next Chapter 6). Every entity weight in the user profiles is multiplied by the corresponding value in Table 5.5 according to the provenance of the entity. This way we can increase the relevance and weight of the interests which were retrieved from more accurate social features. Finally, all the interests weights are then normalised again on a 0 to 1 scale.

	AVG Score	std.dev.
Twitter	3.71	1.34
Facebook	3.48	1.50

Table 5.7.: Average score and standard deviation for interests extracted from Facebook and Twitter only (on a 1 to 5 scale)

The last provenance feature that we analysed as part of our profiling heuristics is the origin or social media source of the interests. We computed the average user scores given to interests extracted from each social media site separately (*i.e.* Facebook and Twitter). This study, as illustrated in Table 5.7, could indicate whether one platform is more suitable for mining user interests than the other. However, from the results of the study we cannot draw a conclusion in this regard. The difference in the score between the two platforms is not significant, with Twitter performing slightly better than Facebook. An hypothesis for this could be that the two platforms allow for similar kinds of actions, despite them being two different types of social media sites. Therefore, in order to evaluate this provenance-related dimension it is necessary to perform an extensive experiment comparing many different types of social media.

5.5. The Need for Privacy and User Profile Management Systems

In the past few years, the growing number of personal information shared on the Web increased awareness regarding privacy and personal data. Recent studies showed that privacy in social networks is a major concern when user profiles are publicly shared, revealing that most users are aware of privacy settings. Most social networks provide privacy settings restricting access to private data to those who are in the user's friends lists (i.e. their "social graph") such as Facebook's privacy preferences. Yet, the studies show that users require more complex privacy settings as current systems do not meet their requirements [Boyd and Hargittai, 2010]. Hence, we propose a platform-independent system that allows end-users to set fine-grained privacy preferences for the creation of privacy-aware faceted user profiles on the Social Web.

Social networks, using non structured data formats, provide minimum privacy settings such as granting privileges to all people belonging to one's social graph to access her

information [Boyd and Hargittai, 2010]. We envisage a social network using structured data which provides users the ability to specify which information can be accessed by specific users who have for instance similar attributes (*e.g.* interests, contact information, etc.). This would make users feel more confident when publishing online their information, especially since they specifically know who can access their information. Although applications are being developed to export user information from closed social networks into structured data such as RDF, the privacy settings are platform dependent such that the privacy settings cannot be reused on other platforms. Moreover, privacy preferences cannot make use of other platform's information, for instance, defining a privacy preference that restricts access to users from one platform and grants users from another platform [Kärger and Siberski, 2010]. Additionally, most social networks have the sole authority of controlling all user's data [Au Yeung et al., 2008]. Therefore, a system that allows users to create fine-grained privacy preferences which can be used by different platforms is required. This system will provide users to be fully in control of who can access their personal information and who can access their published structured data. In this regard, the benefits of using interoperable and standard Semantic Web technologies for managing privacy over personal data are clear [Gandon and Sadeh, 2004].

The system we propose and describe in this section aims at providing a user the necessary tools and options for setting fine-grained privacy preferences on her full private profile which is the result of the aggregation of different distributed profiles from different sources. As displayed in Figure 5.8, the prototype we implemented is composed of two main parts: the *User Profiling* module, and the *Privacy Preference Manager* module – *MyPrivacyManager*. The first part is the component that collects profile data from different social media websites (*e.g.* personal information, activities, interests, etc.), generates specific user profiles for each platform, and then merges them in a global complete user profile. This part has been already described in this chapter and we use the same methodology and implementation for generating aggregated user profiles from multiple sources. The User Profiling module, as presented so far in this thesis, can be connected to any other privacy and profile manager. This is because this module offers profile data described following standard ontologies, such as FOAF and SIOC, as previously described. The second component (*MyPrivacyManager*) described in Section 5.5.1, allows the owner of the full user profile to specify her privacy preferences on the profile. It also manages the requests of other users by asking for the requester's profile information: it replies with a faceted, or filtered, user profile which is the result of the privacy preferences applied to the full profile based on the profile information of the requester.

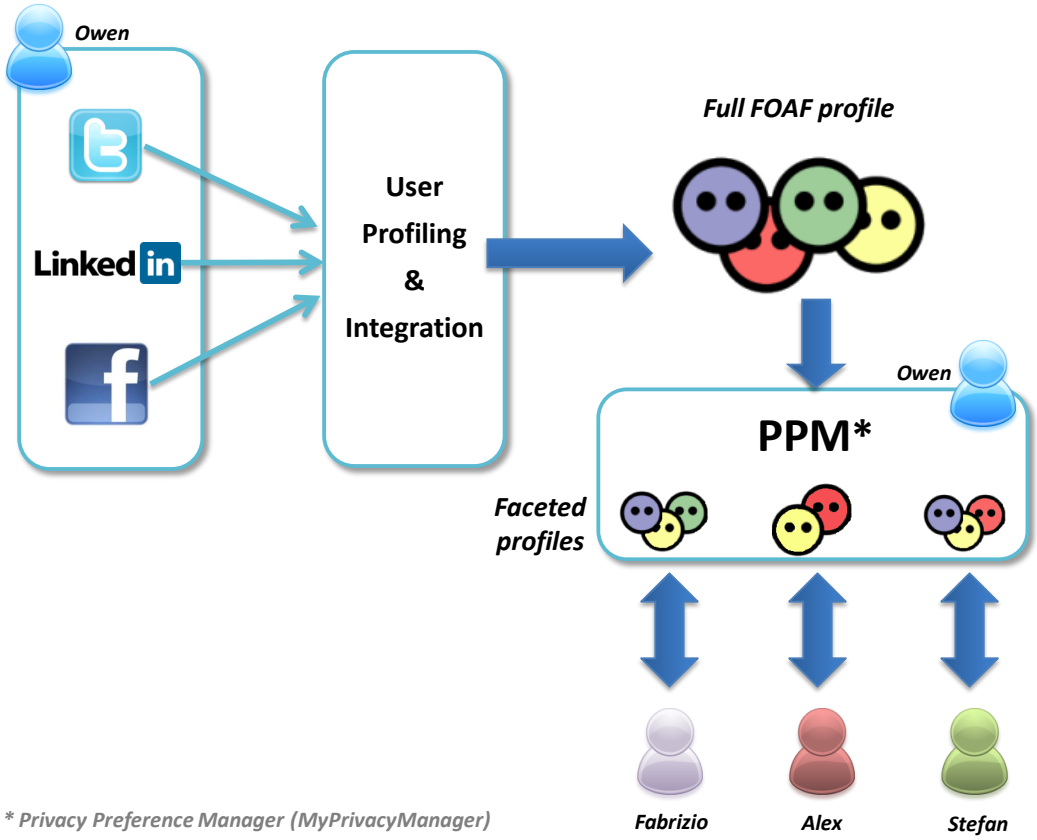

Figure 5.8.: Architecture of the system for user profile and privacy management

5.5.1. A Privacy Preference Manager for Faceted User Profiles

This section presents *MyPrivacyManager*[16], a Web application that serves as a privacy preference manager for the Social Semantic Web. This application is connected to the user profiling application, as the one described earlier in this chapter, in order to provide users with the ability to finely manage their privacy preferences over their distributed Social Web personal data. In the following sections we provide only a brief overview of the software prototype that we implemented. As this is not into the scope of this dissertation, we refer to our publication [Sacco et al., 2012] and other related articles by Sacco et al. for more details on the privacy aspects, the implementation and solutions adopted [Sacco et al., 2011, Sacco and Breslin, 2012]. Relevant related work has been published also in [Villata et al., 2012] where the authors propose a generic semantic access control system for any SPARQL endpoint. Their system, since it is based on

[16] A screencast is available online at: http://vmuss13.deri.ie/faceteduserprofiles/screencast/screencast.html (accessed January 2014)

standard Semantic Web technologies, could also be applied to our user profiling module. Our module would generate semantic user profiles which would be stored in triplestores and could be queried through SPARQL endpoints. Here we briefly describe a possible alternative solution focusing in particular on a personal user profile manager integrating access control capabilities.

5.5.1.1. Architecture

MyPrivacyManager was developed to implement the creation of privacy preferences for RDF data described using the Privacy Preference Ontology (PPO) [Sacco and Breslin, 2012] and make sure the preferences are applied when requesting information to filter requested data. Although *MyPrivacyManager* is designed to work with any Social Semantic Data that consists of Social Web data formatted in RDF (or any other structured format), we will focus on defining privacy preferences for FOAF-based user profiles. With FOAF profiles, our aim is to illustrate how personal information can be filtered based on privacy preferences to generate faceted profiles.

MyPrivacyManager allows users to manage their privacy preferences and also grants access to users' information when requested. The system therefore restricts everything by default and grants access to specific information based on the preferences specified by the users. The architecture provides users to:

(1) Authenticate to their *MyPrivacyManager* instance using the WebID protocol and create privacy preferences based on their FOAF profile; and

(2) Authenticate to third party user's *MyPrivacyManager* instance which automatically requests to view the FOAF profile (of the third party) which is filtered based on privacy preferences.

Figure 5.9 illustrates the *MyPrivacyManager* architecture, which contains:

(1) WebID Authentication: handles user sign-on using the FOAF+SSL protocol (discussed later in this section);

(2) RDF Data Retriever and Parser: retrieves and parses RDF data such as FOAF profiles from WebID URIs;

(3) Creating Privacy Preferences: defines privacy preferences using the PPO ontology;

(4) Requesting and Applying Privacy Preferences: queries the RDF data store to retrieve and enforce privacy preferences;

(5) User Interface: provides users the environment whereby they can create privacy preferences and to view other users' filtered FOAF profiles, hence generating a faceted profile; and

(6) RDF Data store: a RDF data store to store the privacy preferences[17].

The WebID protocol [Story et al., 2009] provides a mechanism whereby users can authenticate using FOAF and SSL certificates. The SSL certificates (which can be self-signed certificates) contain the public key and a URI that points to the location where the FOAF document is stored. Once the user requests to log in *MyPrivacyManager*, the browser prompts the user to select a certificate. The authentication mechanism parses the WebID URI from the certificate and retrieves the FOAF document from its location. The public key in the certificate and the public key in the FOAF file are checked to grant the user access to *MyPrivacyManager* if the public keys match.

MyPrivacyManager uses WebID protocol since it utilises the benefits of URIs where users have a unique identification unlike OpenID[18]. Although OpenID provides a framework where users can log into systems using other system's authentication mechanisms, when users have more than one OpenID account acts as if they identify different persons rather than identifying the same person as how WebID does.

Once the user is authenticated, *MyPrivacyManager* matches the WebID URI with the WebID URI of the owner of that instance. If the owner is signed in, then the interface provides options where the user can create privacy preferences or preview her faceted profile how it appears to specific users. On the other hand, if the user signed in is a requester, then the faceted FOAF profile of the owner of that particular instance is requested. The *Requesting and Applying Privacy Preferences* module is called to filter the FOAF profile according to the privacy preferences specified by the owner of that instance, hence generating a faceted profile.

MyPrivacyManager employs the federated approach whereby everyone has her own instance of *MyPrivacyManager*. As opposed to the majority of Social Web applications which are centralised environments whereby the companies offering such services have the sole authority to control all user's data, this federated approach ensures that everyone

[17]Although ARC2 was used for the implementation of MyPrivacyManager, any RDF store can be used.
[18]OpenID – http://openid.net/ (accessed January 2014)

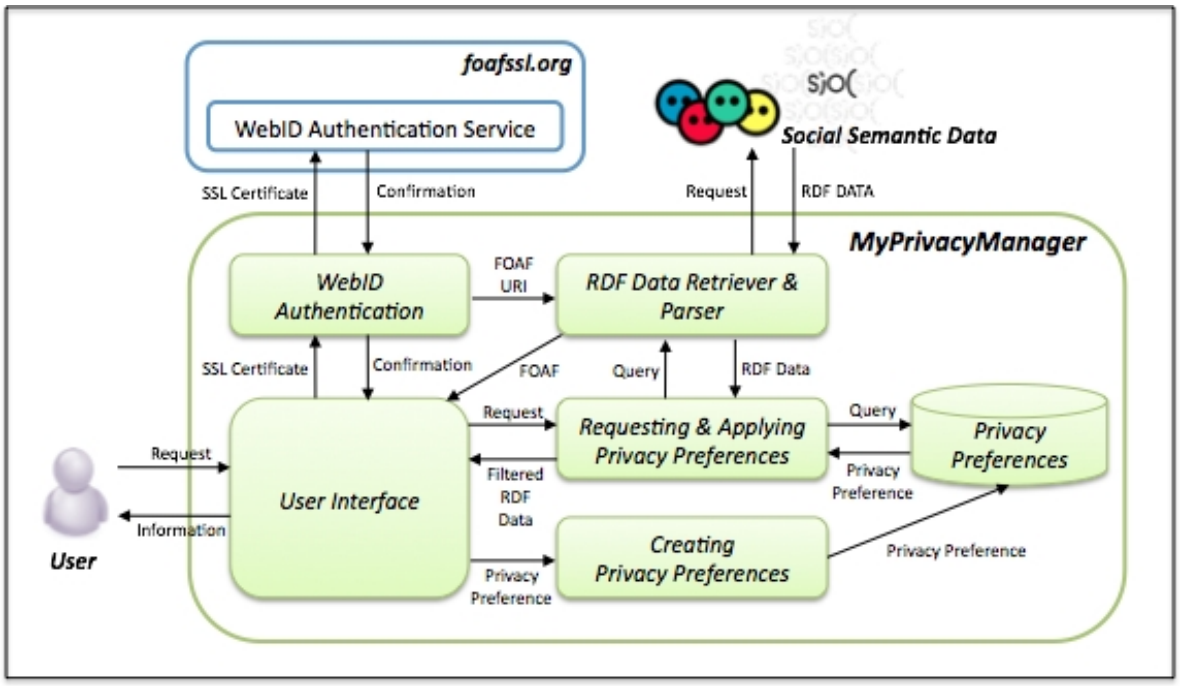

Figure 5.9.: MyPrivacyManager Architecture

is in control of their privacy preferences [Au Yeung et al., 2008]. Moreover, users can deploy their instances of *MyPrivacyManager* on whichever server they prefer. This approach ensures that the FOAF profile and privacy preferences are private since the user becomes the sole authority of her data and nobody can access such data unless he/she is granted access.

5.5.1.2. User Interface

Together with the implementation of *MyPrivacyManager*, we designed an interface to create privacy preferences for the user profiles aggregated from multiple sources. On loading the interface, the system first retrieves and loads all the vocabularies which are used during the creation of the privacy preferences.Once the vocabularies are loaded, the system retrieves the full FOAF-based user profile (generated from Twitter, LinkedIn and Facebook) from the WebID URI contained within the SSL certificate. The interface then displays (1) the profile attributes which the user can specify what to share in the first column and (2) other attributes (extracted from the user profile) in the second column for the user to specify who can access the specific shared information; — as illustrated in Figure 5.10.

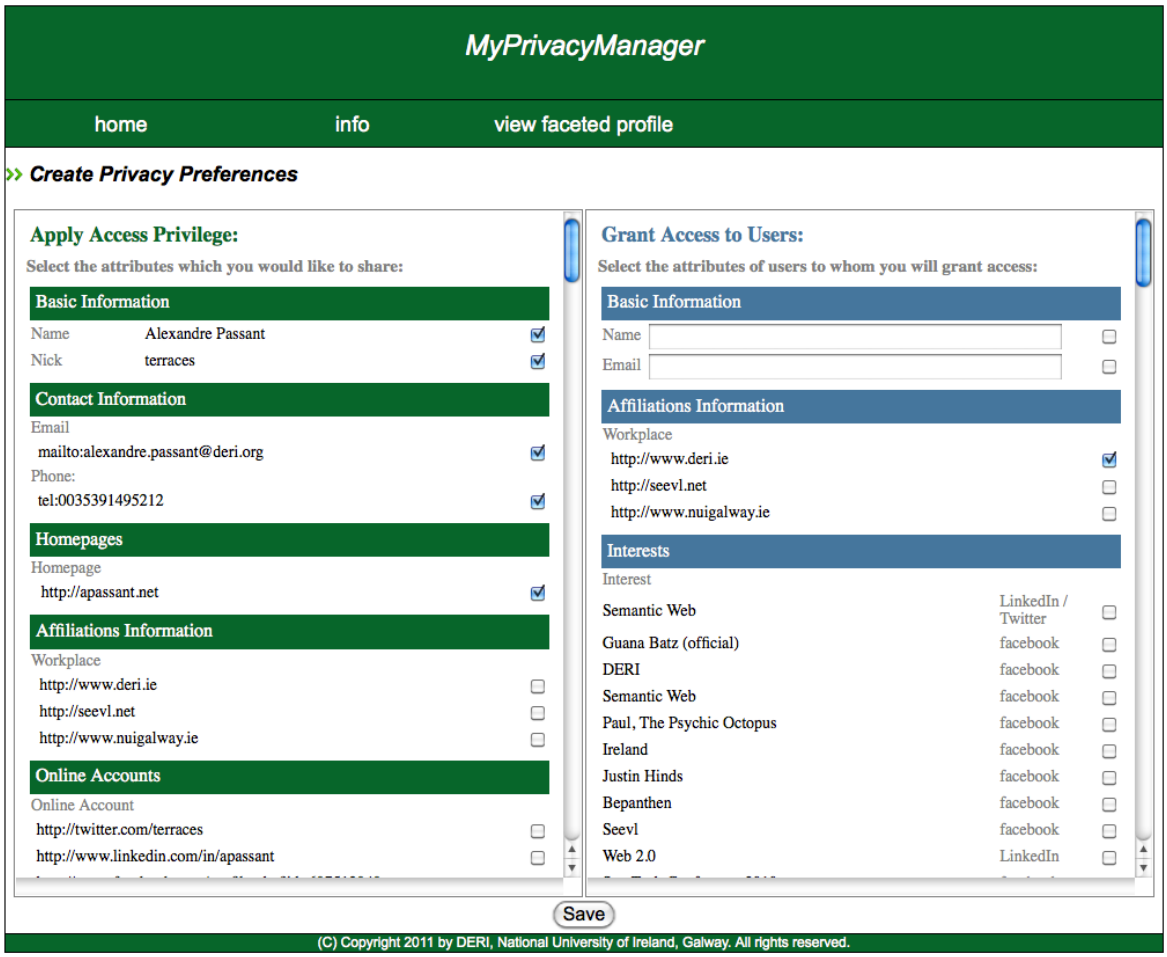

Figure 5.10.: The interface for creating privacy preferences in MyPrivacyManager

Before the development of the interface, we conducted a preliminary user study (included in Appendix A.1) to motivate and refine the design of the interface. The outcome of the study clearly shows that users want to specify different privacy preferences for different groups of their profile information. Therefore, our system provides profile attributes which the user can share classified as follows:

(1) Basic Information, consisting of the name, age, birthday and gender;

(2) Contact Information, consisting of email and phone number;

(3) Homepages;

(4) Affiliations, consisting of the website of the user's work place;

(5) Online Accounts, such as Twitter LinkedIn and Facebook user pages;

(6) Education, that contains the user's educational achievements and from which institute such achievements where obtained;

(7) Experiences, consisting of job experiences which include job title and organisation; and

(8) Interests, which contain a list of user interests ranked according to the calculated weight of each interest.

Moreover, the user study in Appendix A.1 demonstrates which attributes the users prefer to specify and to whom they want to share their information with. This study shows that users prefer to select specific users from a contacts list. Since a considerate number of users have selected that they would require sharing information without knowing who the person is, we opted to not provide any user contact lists but provide users to specify the attributes of whom they want to share information with. Our aim is to study whether users are satisfied with our approach which provides sharing information to a greater (or less) audience without knowing 'a priori' who the person is and without having the user maintain user lists. For this reason we conducted a user evaluation for the interface and the whole system (included in Appendix A.2) which shows that users accepted our approach and were satisfied how the system granted access. The attributes the user can select to whom to share information are extracted from the FOAF profile and provided by the system. They are categorised as follow:

(1) Basic Information containing fields to insert the name and email address of specific users;

(2) Affiliations to share information with work colleagues; and

(3) Interests to share information with users having the same interests.

Once the user selects which information to share and to whom, he/she clicks on the save button for the system to generate automatically the privacy preference. Hence, the application generates automatically the restrictions, conditions and access space query automatically based on what the user selected.

We envision different solutions that could be adopted for the update, or the automated creation, of privacy preferences for new interests. When new interests are introduced into the user profile, specific preferences can be automatically adopted according to the provenance, or the type, of the interests. For example, specific rules can be automatically triggered for interests which have been extracted from a particular social

media source, or are the result of specific actions, or belong to certain topics. This is a very interesting research topic that need to be investigated in the near future, however it goes out of the scope of this dissertation.

With this work we introduced a system providing users full control over their personal user profile allowing them to define and show different *facets* of their profile based on fine-grained privacy preferences. We described the architecture of the user profiling module of the system and the methodology proposed for the aggregation of different user profiles on the Social Web. Moreover, we provided a brief overview on the structure of a privacy preference manager - MyPrivacyManager, which allows the specification of the privacy preferences on the profile data. Additionally it also provides users to verify their faceted profiles as visible by other users. The architecture proposed is applicable to any kind of site on the Social Web, and MyPrivacyManager is also platform independent. We argue that more research on this topic is necessary and systems such as the one proposed here need to be developed further. Related similar research is available in the state of the art [Villata et al., 2012] and demonstrates the increasing interest on solutions to the user profiling and privacy challenge. However, to the best of our knowledge, we currently do not see any popular system available and used by many users on the Web. A system that allows users to manage their own distributed user profile and protect it with privacy preferences. The proposed research opens another very important aspect of user profiling on the Social Web that should not be ignored when dealing with solutions for the management of personal user data.

5.6. Conclusions

In this chapter, we described the core of our methodology for the automatic creation and aggregation of interoperable and multi-domain user profiles of interests. In particular, we focused on two essential steps of our methodology: *aggregating* and *mining* user interests extracted from social media data along with its related provenance information (as described in the previous chapters). We demonstrate the efficiency and accuracy of entity-based user profiles of interests as compared to traditional tag-based techniques. The potential of the entities of interest connected to the Web of Data is shown and their efficiency for the aggregation task and semantic enrichment is evaluated. We evaluate the effect of different provenance-based dimensions and heuristics on mining and ranking entities of interest in order to increase the accuracy of the profiles for the users. In

this regard, two user studies (with 21 and 27 volunteers respectively) conducted using Facebook and Twitter user accounts, are detailed in the chapter. The outcome of the study provides an insight on the most accurate features of users' social data for profiling interests and on the impact on accuracy of entity-based versus category-based profiles, and different time decay functions. We conclude the chapter illustrating the importance of privacy in our research and describe a management system for privacy preferences on user profile data.

Chapter 6

Semantic Enrichment of User Profiles of Interests for Personalisation

6.1. Introduction

Extracting and representing user interests on the Social Web is becoming an essential part of the Web for personalisation and recommendations (see Chapter 1). Such personalisation is required in order to provide an adaptive Web to users, where content fits their preferences, background and current interests, making the Web more social and relevant. Profiling interests of online communities is also a challenging research field that could provide insights on the evolution and propagation of knowledge and culture on the Web. So far, in this thesis, we have explored how to mine, aggregate and represent user interests from different sources on the Social Web. In particular, semantic representation of the concepts of interest revealed to be essential for selecting and filtering user interests according to several measures and the particular use case. Different use cases, or personalisation tasks (such as recommendations or user adaptive interfaces), require distinct types of concepts of interest and therefore specific profiling strategies.

As anticipated in Section 5.2.3, we use the Web of Data, namely DBpedia, not only to link to its entities but also to extract related categories for concept expansion. We can analyse the structure of the concepts graph in order to understand the relevance of entities and/or categories for representing user interests. Representing interests using DBpedia resources has two main advantages: it integrates the user profiles with the Linked Data cloud, and it provides a larger and "fresher" set of terms as compared to

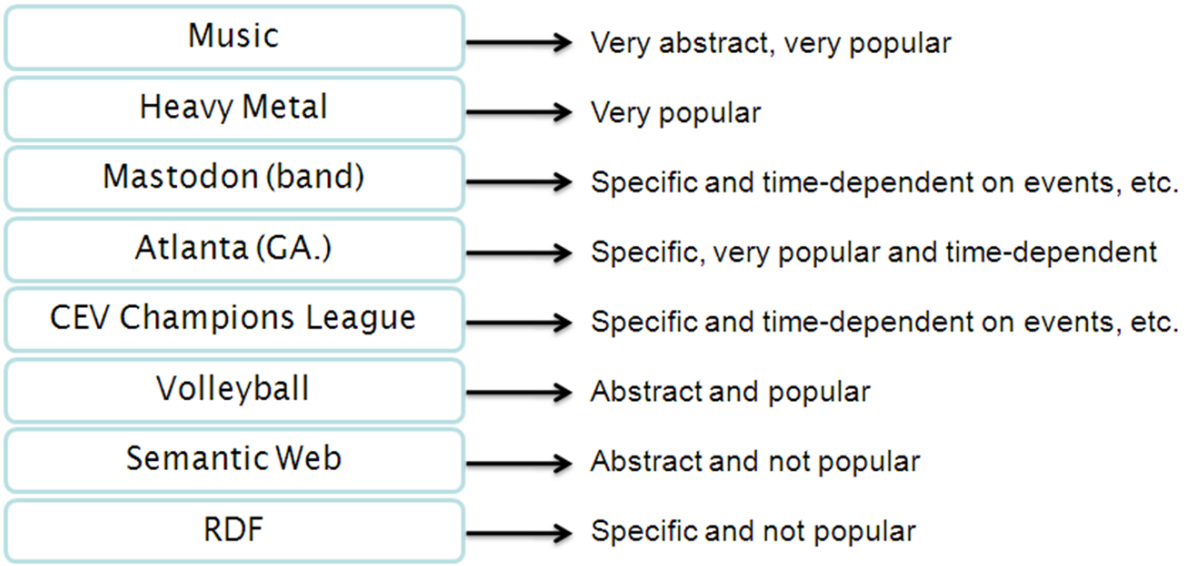

Figure 6.1.: Example of different dimensions of entities of interest in a user profile. We need a deeper understanding of the semantics and pragmatics of the entities.

any other knowledge base. By exploring the Linked Data graph we can relate information to the original concepts of interest of any of our entity-based profiles and enrich their semantics. Therefore, in this chapter we describe our methodology for semantic enrichment and characterisation of concepts of interest. We employ the Web of Data and the Social Web for the enrichment and evaluate the impact of our measures on selected personalisation scenarios.

In Section 6.2 we discuss the limitations of entity-based user profiles of interests (such as the ones we described and evaluated in the previous Chapter 5), as they are often missing the semantics of the entities in terms of: (i) categorisation, (ii) popularity and temporal dynamics of the interests on the Social Web and (iii) abstractness of the entities in the real world. State of the art techniques to compute these values are using specific knowledge bases or taxonomies and need to analyse the dynamics of the entities over a period of time. Hence, we propose a real-time, computationally inexpensive, domain independent model for concepts of interest composed of: popularity, temporal dynamics and specificity.

Additionally, we describe how to deploy user profiles on practical personalisation use cases. The impact of our profiling methodology on a personalisation system for real-time Social Web streams is evaluated in Section 6.3. We propose a methodology and a set of heuristics to filter any public and large social stream of short textual messages and

personalise it in real-time according to automatically updated user profiles of interests. We describe the theoretical background and the implementation of "SPOTS" (Semantic Personalisation Of the Twitter Stream) a system offering real-time personalisation of the public Twitter stream. SPOTS aims at recommending interesting tweets to users according to (i) their implicitly/explicitly shared preferences on the Social Web, (ii) additional information extracted from the Web of Data in real-time and (iii) specific informativeness measures. We provide a user-centric evaluation of the system by comparing it to the official Twitter "Discover"[1] service and give insights about the scalability and real time nature of the implementation.

6.2. Linked Data and Social Web for User Profiles of Interests

6.2.1. Enriching User Interests Using Linked Data

A very important phase of our methodology for user profiling is the semantic enrichment of concepts of interest. With the term "semantic enrichment" we describe the process of connecting entities on the Linked Data cloud and using the potential of this knowledge graph for expanding our information about the represented entities (Figure 6.2). This knowledge expansion can have several forms and goals: from the computation of semantic relatedness among concepts, to the discovery of several properties connected to the entities of interest, to their categorisation, etc.

In particular, we propose the use of DBpedia to represent the interests of the users (as described in Section 5.2.3). We decided to use DBpedia because of our particular use case (domain-agnostic) and to facilitate our implementation. However, with our methodology we could have used the entire Web of Data, or any other knowledge base, exposed following Linked Data principles. As previously described, DBpedia is the semantic representation of Wikipedia and it has become one of the most important and interlinked datasets on the Web of Data. Compared to traditional taxonomies or lexical databases (e.g. WordNet) it provides a larger and "fresher" set of terms, continuously updated by the Wikipedia community and integrated into the Web of Data. The benefits of using DBpedia for this purpose are described in Section 5.2.3 as well as in [Ponzetto and Strube, 2007]. In particular, in Chapter 5, we already described

[1] https://support.twitter.com/groups/53-discover (accessed January 2014)

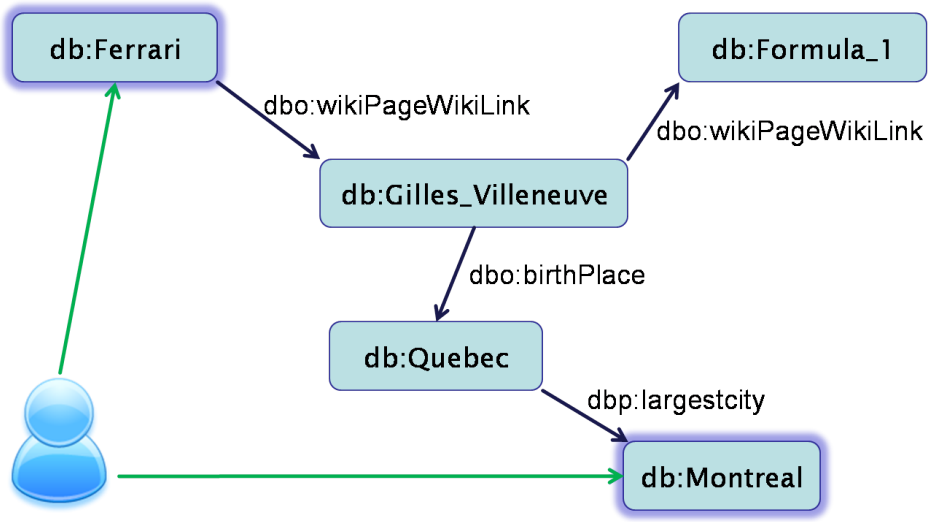

Figure 6.2.: Example of semantic relatedness of two concepts on DBpedia showing the potential of Linked Data for user profiling. Here in the example, "Ferrari" and "Montreal" were already in a user profile and apparently disconnected, but on DBpedia they revealed to be closely related.

how we use DBpedia not only to link to its entities but also to extract related categories for concept expansion. In Section 5.3.3 we detailed how we analyse the structure of the categories graph in order to understand the relevance of a category for representing a user interest. The outcome of our study on user profiles consisting of DBpedia categories shows that this type of user profiles are slightly less accurate than the ones consisting only of DBpedia resources. However, categories provide an expansion of the original resource-based profiles in the number of concepts of interests available. According to our methodology, category-based user profiles are 7 times richer than resource-based profiles.

Another interesting aspect of semantic enrichment using Linked Data is the straightforward categorisation of the entities of interest according to their type. Every resource on DBpedia is an instance of a class which identifies its type in an ontology. By analysing the type of each entity (on DBpedia expressed using not only the `rdf:type` property but also a specific `dbpedia-owl:type` property defined in the DBpedia ontology) expressed by a simple property in RDF, we are able to identify the type of the interests as well. For example we can identify interests belonging to *Music* types, or *Populated Places*, *Movies*, *People*, etc. This feature is very useful for grouping and filtering the interests of a user profile, an operation that is usually performed in specific personalisation tasks.

6.2.2. Concepts' Abstraction

In this section we describe a methodology for efficiently computing the level of specificity, or term abstractness, of a particular real-world entity or concept that is uniquely representable on the Web. As for *specificity*, we define it as the level of abstraction that an entity has in a common conceptual schema shared by humans. Human knowledge can be organised in taxonomies where concepts and instances (in general, entities) are categorised and related to each other with broader/narrower relations. These relations for instance reflect and determine the specificity of the entities in a hierarchical classification system. Entities positioned at high positions in a taxonomy are considered less specific (or broader, or more generic) than entities positioned in lower positions of the taxonomy (hence closer to the leaves of the hierarchy). As an example, according to our definition, the entity representing *Alternative Rock Music* is more specific (or has a higher degree of specificity, and lower abstractness) than the entity *Music*. In this work we present a novel approach to automatically determine the specificity of entities and hence to improve personalisation on the Social Web.

This measure expresses how abstract an entity is in a common conceptual schema and does not refer to the popularity of the term. A real-world entity can at the same time be very generic but not very popular in Social Media systems (e.g. "Classical Music") or can also be both very specific and very popular (e.g. a Pop/Rock song of the moment). This is why for characterising and ranking the relevance of the entities of interest we need to combine this dimension with popularity features (described in this Section 6.2). We note also that these features are user-independent and are only computed using information retrieved from the Social and Semantic Web. In order to be applied to user adaptive systems for recommendations or personalisation they need to be combined with user-based relevance measures. For instance, once the entities of interest in a user profile are ranked according to a relevance score for the user, they can be re-ranked also using these particular features. These features can be tuned according to the use case and the system they are implemented on (e.g. for filtering Twitter we should prioritise specific, popular and trendy interests).

Several state of the art approaches, in order to compute specificity, utilise a taxonomy of concepts which are categorised and organised in a hierarchical structure. The more the entities are categorised in a position of the hierarchy close to the top or the root the more they are considered generic. Hence the specificity of the entities increases when going from the root to the leaves of the categorisation tree. This approach works well in many

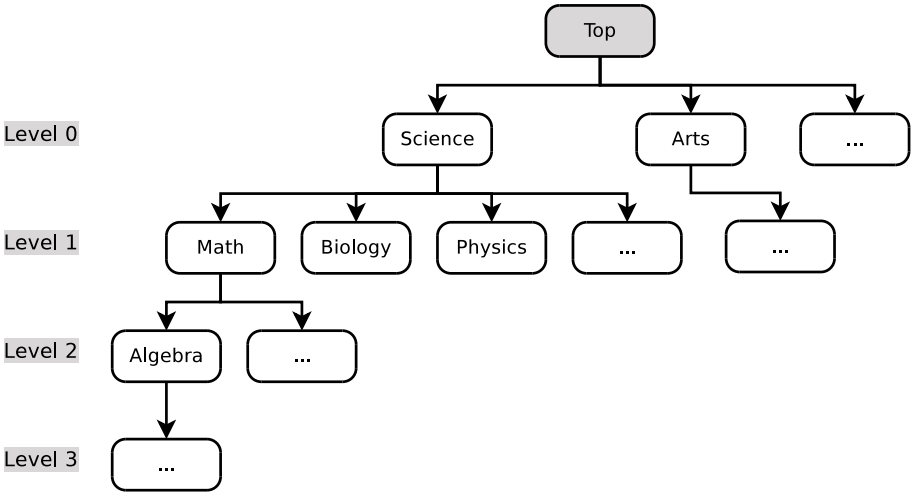

Figure 6.3.: Example of a taxonomy: a portion of the DMOZ taxonomy used as a comparison for evaluation purposes

situations and is clearly justified by the fact that a taxonomy is by definition organised by supertype-subtype relationships, also called *generalisation-specialisation* relationships. Moreover, this approach works only with taxonomies organised in a tree-like structure, hence not for graphs which could present cycles. Therefore, the problem comes when the knowledge base, that has to be used for a particular use case, needs to be (i) continuously updated with the evolution of the entities and the events in the real-world, (ii) organised in a tree structure, (iii) universal (not restricted to a particular domain) and (iv) suitable for real-time computation. Many large and available knowledge bases have been used in research for this purpose such as Wikipedia, DMOZ, WordNet, OpenCyc[2], etc. but they do not satisfy all the aforementioned requirements. Wikipedia/DBpedia for example, is continuously updated and very large but its category structure is not a hierarchy but a graph [Ponzetto and Strube, 2007]. DMOZ, Wordnet and OpenCyc present a hierarchical structure but are not continuously updated by a large collaborative mass of users who keep the knowledge base up-to-date.

Following these requirements we decided then to directly use the potential offered by the Web of Data as background knowledge. Thus, instead of using measures for hierarchical structures we use graph measures on the Web of Data graph leveraging the Linked Data principles. Because the entities and concepts are represented on the Linked Data cloud as nodes of a network, common network properties can be measured. In particular, we can consider the Linked Data network as a directed labelled graph.

[2]DMOZ: www.dmoz.org, Wordnet: wordnet.princeton.edu, OpenCyc: www.cyc.com/platform/opencyc (accessed January 2014)

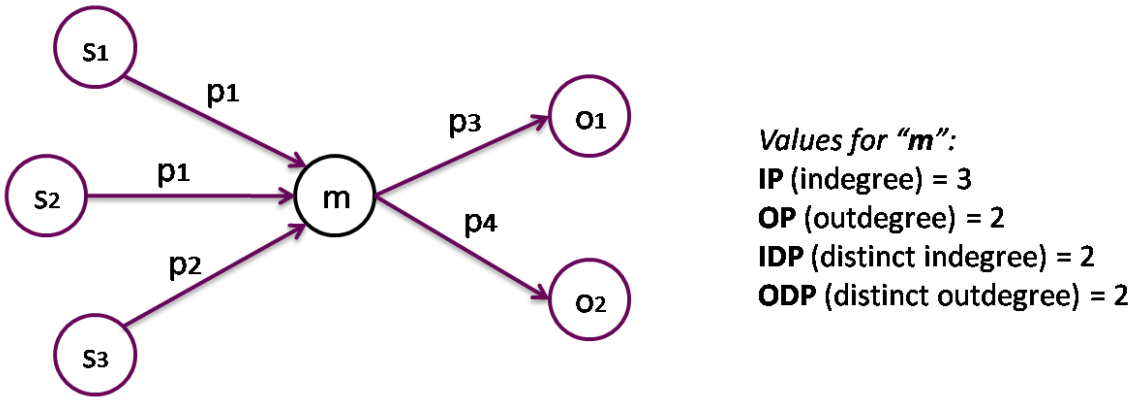

Figure 6.4.: Notation used for the specificity measure and example.

In this work, we use the notion of *indegree* and rename it with the acronym "**IP**" **(Incoming Predicates)** as we are working with the Linked Data labelled graph, and we do the same with *outdegree* with "**OP**" **(Outgoing Predicates)**. Additionally, we also use other two measures that we call *"distinct outdegree"* and *"distinct indegree"* of a vertex v in a graph, that we define respectively as the number of distinct edges with tail v and the number of distinct edges with head v. Here we will call these two measures "**ODP**" **(Outgoing Distinct Predicates)** and "**IDP**" **(Incoming Distinct Predicates)**. They basically represent the number of distinct predicates that are connected to a subject or an object. Hence, the distinction is made in terms of type of the edge or property, not considering the objects/subjects connected. Nodes with many different types of outgoing properties connected have high values of ODP and, similarly, nodes with many different types of incoming properties connected have high values of IDP.

Analysing the predicates connecting entities on the LOD cloud (and in particular on DBpedia) we noticed that very specific entities have many different types of outgoing predicates compared to the incoming ones. On the other hand for generic ones the two numbers are usually comparable or the distinct incoming predicates dominate over the number of the outgoing ones. From this observation then the hypothesis is that the ratio between Incoming Distinct Predicates (IDP) and Outgoing Distinct Predicates (ODP) characterises the specificity of entities. Thus, we formulate *our measure for specificity*:

$$\mathbf{DRR}(DistinctRelationsRatio) = \frac{IDP}{ODP} \tag{6.1}$$

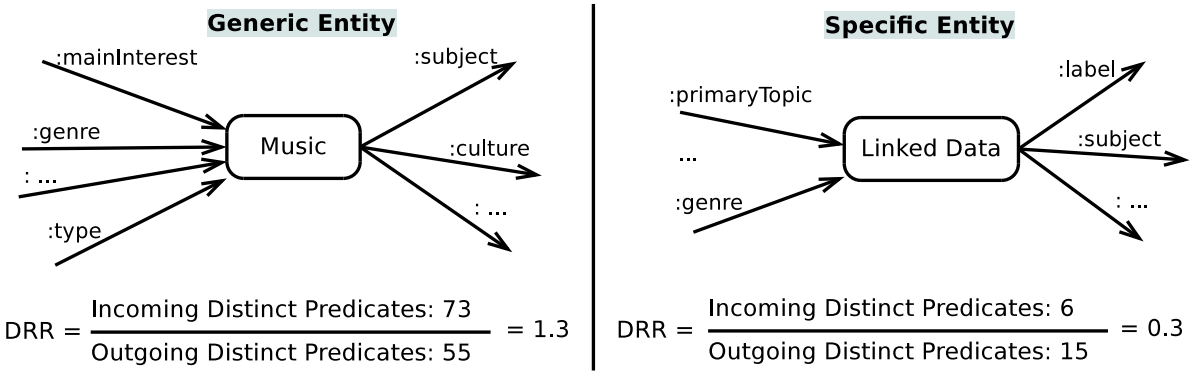

Figure 6.5.: Example of the DRR measure with two entities: one generic and one specific.

This measure relies on the orientation of the predicates on the LOD. Therefore, the same measure would not work correctly if we would consider also the inverse properties of the existing LOD predicates. The orientation is important as it reflects the natural orientation given by humans to properties when creating Linked Data datasets. Moreover, it is interesting to note that *Literals* cannot be used as subjects according to the Semantic Web principles. Hence, literals play a crucial role for this measure as well. Additionally, as suggested also in [Theoharis et al., 2008] where the authors show that the position of classes in subsumption graphs is influenced by their in/outdegree, the measure we propose for specificity can be logically justified. It is reasonable to think that very specific entities have a high variety of predicates pointing toward many other entities and few other entities pointing at them with different predicates. A very specific entity is also "mentioned" always in one particular sense and context by other subjects so it can also have many incoming links but they will all be of the same kind (hence low IDP). The other very important advantage of the IDP/ODP measure is that it is simple and can be easily computed without intensive use of computational resources. In fact, the datasets of entities on the LOD cloud are almost all of them indexed and can be queried on the Sindice project SPARQL endpoint[3]. Using a simple SPARQL query it is possible to interrogate the entities represented in RDF on the Web of Data and for instance get their in/outdegrees almost instantly. We tested these queries on the Sindice endpoint and we note that for every such query we made, it always returned a (non-empty) result in less than one second.

We compared our DRR measure with other similar measures such as: IP/OP, IP+OP, IP and different state of the art approaches in a preliminary evaluation. In this section, we briefly describe the main experiments conducted in order to evaluate the performance of our DRR measure compared to a gold standard given by 5 human evaluators

[3] http://sindice.com/ (accessed January 2014)

and aimed at verifying the validity of our approach for a binary classification task (i.e. for the automated classification of entities as *Generic* or *Specific*). For more details on the DRR measure for specificity we invite the reader to consult Appendix B (and our publication [Orlandi et al., 2013]) where we included a description of the complete set of experiments conducted for evaluation purposes. Here, the methodologies evaluated are compared against our gold standard, the user manual classification. We had an evaluation dataset of 160 random DBpedia entities from user profiles of interests and we performed the classifications with different methodologies. Five evaluators created the gold standard by manually categorising the entities into two categories according to their level of specificity. We computed the Fleiss' generalised Kappa coefficient for 160 subjects, 5 raters and 2 categories and we obtained $K = 0.61$, which is an indication of moderate/substantial agreement (according to Rietveld and van Hout (1993) [Eugenio, 2000]). We manually classified the entities following a state of the art method using the DMOZ taxonomy and according to the position of the entities in the DMOZ hierarchical tree (as depicted in Figure 6.3). Afterwards, the precision of the DMOZ classification, the DRR, and other Linked Data-based measures (IP/OP, IP+OP, IP) have been computed against the manual classification performed by the 5 human evaluators. For *precision* here we intend the number of entities classified in the same way by the two methods over the total number of entities of the dataset. As we can see in Table 6.1 the DRR measure and the DMOZ classification have similar performance compared to the manual classification. For around 84% of the entities the two strategies classified the entities in the same way as the human evaluators. All the other LOD-based measures perform clearly worse in this classification task. According to these results our automatic measure has comparable performance with state of the art approaches such as those using a taxonomy like DMOZ as a background knowledge. The clear advantages in using Linked Data is that the background knowledge is extended on a Web scale, it is always updated with the quick evolution of the Social Web, it does not need to be pre-processed or stored and simple measures like the DRR can be computed in real-time. For more details and a more extensive evaluation on the DRR measure for specificity we invite the reader to consult Appendix B (and our publication [Orlandi et al., 2013]). The outcome of this work confirms the DRR as a good approximation for measuring the specificity of an entity.

DMOZ	DRR	IP/OP	IP+OP	IP
0.839	**0.841**	0.700	0.700	0.725

Table 6.1.: Specificity evaluation: precision of the different methods compared to the manual user classification.

6.2.3. "Trends" and Temporal Aspects of Concepts

The dynamics of the frequency of mentions of entities on the Social Web over time provide accurate understanding on the popularity of terms, their evolution and their trend. Understanding these dynamics can be very useful for instance in recommendation systems. It would be possible to discard entities of interest from recommendations if they were popular only at a very specific point of time in the past (as they might be related to a specific past event), or to increase the weight and relevance of currently popular entities. This measurement is particularly useful to characterise trendy entities which are very time-dependant and entities showing stable dynamics of popularity over time. For this measure we propose the usage of the Wikipedia page views. The number of Wikipedia page views for every day, every year, are publicly available and accessible through the Wikipedia MediaWiki API. This source of information provides an effective way of detecting the interests of the users on entities over time. In our experiments we use mean and standard deviation of the number of views for the past 30 days to distinguish concepts that are steadily popular over time or present relevant fluctuations in their dynamics. Following empirical experimentation, we define as *"stable"* a resource that has low standard deviation value for the page views of the last recent days (and "unstable" otherwise). Additionally, we define as *"trendy"* an entity that shows a clear increase of the number of page views only in the last recent days. In this case we use a simple method based on linear regression of the page views. In both cases a threshold has to be defined through empirical studies in order to perform the categorisation. Similarly to the other measures, this methodology is not computationally intensive, it is simple, and it is based on a large and continuously updated knowledge base.

In Figure 6.6 we show a diagram of the trend of a popular racing cyclist extracted from a freely accessible Web service. As we can see, the spike in the page views corresponds to a particular event which has been widely discussed in the news.

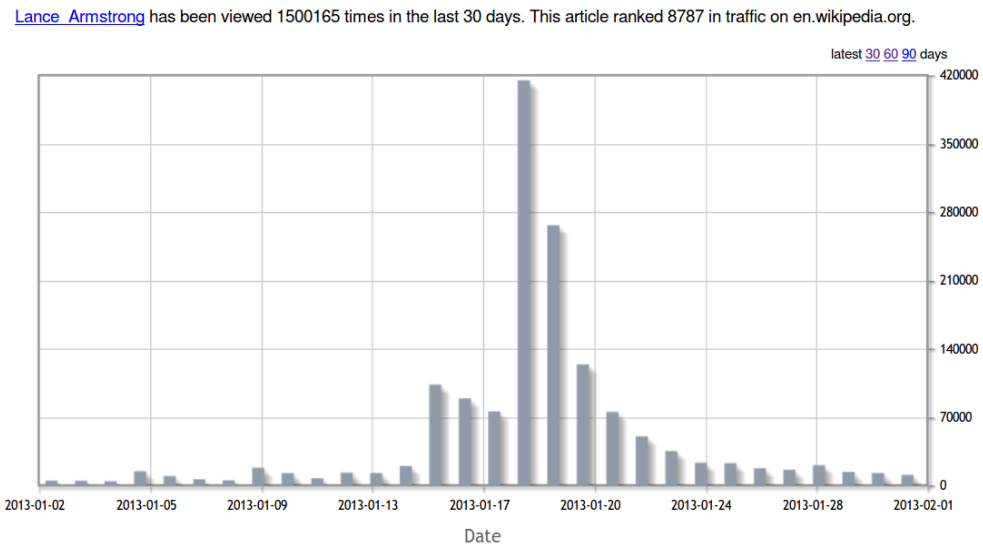

Figure 6.6.: Diagram of the Wikipedia page views for the article "Lance Armstrong" (on January 2013), from: http://stats.grok.se

6.2.4. Popularity of Concepts

Popularity expresses how much the entity is well known, shared or interesting to the majority of the people on the Web. This can be easily measured by looking at the frequency the entity is being mentioned on social media systems and for this particular case we can use Twitter as the system where we evaluate our methodology. Naturally, the fact that we employ Twitter as the only source for this measure can provide biased results. However the large number of users on this microblogging platform and the extensive studies conducted on this source of information [Miller, 2011] demonstrate that the popularity of the entities being spoken about in the real-world is very close to their popularity on Twitter. What we suggest is a straightforward approach that utilizes the Twitter Search API to monitor the frequency an entity has been mentioned by users in a recent time frame. This is done with specific tools for named entity recognition and disambiguation on the resulting tweets from an initial search query in order to filter out ambiguous results. The result of this method is in our case an application that, given a DBpedia resource in input, returns a number representing the tweets per second being generated on Twitter about that entity (high numbers of tweets per second mean high popularity and vice versa). Another advantage of using this measurement on Twitter is that it allows a fast and real-time computation of the popularity, which is very important in many personalisation scenarios (see Section 6.3.1.1). While this measure provides an instant picture of an entity at that specific point of time (a snapshot), it

does not consider the temporal evolution of the popularity over time. Indeed, an entity or concept can be very popular at one specific point of time but not popular considering a longer period of time or vice versa. For this reason we propose the combination of this measurement with the other feature that considers the temporal dynamics of popularity, as introduced in the previous section.

6.3. User Profiles of Interests for Social Web Personalisation

6.3.1. Real-time Personalisation of a Social Web Stream

The vast success of the Social Web over the past two decades has changed the way millions of people communicate, interact and consume information. Streams of social actions (status messages, media object shares, user preferences, comments, etc.) performed on the Web are constantly growing at an unprecedented rate. The wide adoption of microblogging services like Twitter and their real-time nature introduced new possibilities and challenges [Kwak et al., 2010]. At the same time the way we consume information about our own interests has also been revolutionised. The need for fresh, real-time updates and the importance of opinions expressed by online communities are becoming more and more preponderant.

Twitter's non-reciprocative paradigm has encouraged users to follow people based on overlapping interests [Chen et al., 2010]. However, this has lead to two main issues:

1. *Information Overload*: where users' intents (overlapping interests) are ignored as they receive all the tweets of their followees;
2. *Coverage*: where users can evidently miss relevant interesting information from other expert users who are unknown to them (not followees).

Tackling information overload has gained prominence by the demand for applications and commercial Web services[4] offering temporary "muting" capabilities to Twitter users, allowing them to reduce the overload of messages on popular but overplayed or uninteresting topics. However, in [Bernstein et al., 2010] the authors argue that the only

[4]Services such as: http://twitterrific.com/ios, http://muuter.com/, http://mutetweets.com/ (accessed January 2014)

possibility for active microblogging users to keep control over their streams is to constantly refine their lists of followees.

Twitter users receive more than a thousand tweets each day from their followees where only a portion are of their interest. Studies, such as in [Ehrlich and Shami, 2010] and [Bernstein et al., 2010], show the negative effect of the increasing volume of tweets on the Twitter users: they perceive their own feeds as overwhelming. Also, the non-reciprocative paradigm of Twitter has encouraged users to follow other users based on overlapping interests. This is clearly due to the user-centric nature of the Twitter social network, and most of the Social Web streams in general, where users follow other users' updates and not updates about some specific topics and/or interests [Chen et al., 2010]. The rate of unsubscriptions (unfollow) performed by Twitter users is also quite relevant: according to [Kivran-Swaine et al., 2011] on average, a single Twitter user loses about 39% of their followers over a nine months period. Primarily, users cease following those who post "many tweets within a short time" or "create tweets about uninteresting topics" [Kwak et al., 2011]. As a matter of fact, currently the only possibility for active microblogging users to keep control over their streams is to constantly refine their lists of followees [Bernstein et al., 2010]. A user study presented in this paper by Bernstein et al. confirms this problem and supports our work.

In this section we propose a system to filter and recommend tweets from the public Twitter stream, based on user profiles (interests) generated by mining their activity of multiple social networks. The approach is flexible and can be adapted to other Social Web streams. Personalisation is made possible thanks to automatically generated profiles of interests mined from the users' actions on the Social Web, as already described in the previous chapters. Here we refer to the system with the acronym SPOTS (Semantic Personalisation Of the Twitter Stream). We apply our methodology to Twitter as it is currently the most popular microblogging service, however our solutions could be applied to any similar Social Web stream. The choice of personalising the full public stream of posts is motivated by the need of active Twitter readers to focus on the topics of the posts instead of continuously refining their contacts lists. Moreover, this solution provides users the ability to receive interesting posts from the whole Twitter community and not only from a selected subset of users making the discovery of both interesting *tweets* and *users* broader. Personalising a large Social Web stream, such as the public Twitter one, and avoiding information overload at the same time is a non-trivial task. The approach is built on top of the profiling methodology described so far in this dissertation (see also our publication [Orlandi et al., 2012] and other relevant state

of the art contributions [Abel et al., 2011c, Abel et al., 2012, Chen et al., 2010]). Previous related work focused on personalisation on microblogs in particular scenarios. In [Chen et al., 2010] the authors provide topic filtering and personalisation techniques for a restricted subset of the Twitter stream, in particular only for the stream generated by the user's followees or for trendy topics. While in [Abel et al., 2011c] a recommendation system for Twitter is presented but it is not targeted at real-time recommendations and it is restricted to tweets linked to online news articles or tweets about incidents or similar particular events [Abel et al., 2012]. The main challenge in our work is to be able to prioritize the appropriate interests of each user and recommend only messages related to those few selected topics. In addition, there are the challenges of filtering a huge stream of tweets[5] in real-time by measuring their level of informativeness and also the limiting condition that all the computation required for the recommendations has to be done in real-time[6].

To summarise, in the following sections we illustrate:

- novel measures for semantically enriching and characterising concepts of interest and their implementation in a personalisation system;
- the implementation of a distributed real-time recommendation system for microposts of a large social stream such as Twitter;
- specific measures for filtering noisy and non-informative tweets from the public Twitter stream in real-time;
- a user-based evaluation of the performance of our recommendation system compared to a similar system offered by Twitter itself ("Discover").

Twitter Discover[7] (also known as the "Discover Tab") is a service officially provided by Twitter that recommends interesting tweets to its users. The service however does not provide real-time recommendations but it suggests popular tweets from a few minutes old to a couple of days old. More details about it and an overview of our entire personalisation system are presented in Section 6.3.2.

[5] http://scoop.intel.com/what-happens-in-an-internet-minute/ (accessed January 2014)
[6] "Real-time" is the short interval between the publication and recommendation of the tweet.
[7] https://support.twitter.com/groups/53-discover (accessed January 2014)

6.3.1.1. Scenario

In this section we introduce a possible scenario in order to clarify the goals of our research and introduce requirements for SPOTS. Let us consider the following scenario: Alice is a Twitter and Facebook user. On Facebook she demonstrates through some social actions such as posts or comments that she is interested in updates about Dublin (the city) as she will soon spend a week there for some business meetings. Since she is living in London and she never went to Dublin before, she never expressed any interest on that city and she will probably hold that interest only for a short period of time. The aim of a system such as SPOTS is to capture that interest and its relevance to the user from her Social Web activities across heterogeneous social networking platforms and, consequently, to provide interesting updates to the user from the public Twitter stream in real-time. In this way Alice, as a SPOTS user, would be able to receive interesting updates about a musical event being held in Dublin in the following week.

As we can see from the proposed example a few aspects of this work are crucial:

- the ability to "spot" a few entities of interest from Social Web actions and rank them according to their relevance for the user,
- the capability to dynamically update this ranked list of interests,
- the real-time nature of the implemented algorithms and
- the necessary informativeness metrics for filtering a large number of posts about interesting selected topics and hence avoid overload of messages.

6.3.2. Real-Time Nature and Architecture of a Recommender System

We propose a software architecture and introduce new methodologies for personalising and filtering Social Web streams of messages in real-time and for dealing with the problem of information overload generated by most of the popular social media websites. In particular, we implement our methodology on the public stream of Twitter. Further, we automatically extract user interests by combining user activity analysis on two different social media websites: Twitter and Facebook. A semantic distributed pub/sub architecture is responsible for analysing and annotating the tweets from the social stream and deliver them in real-time to the users according to their interests.

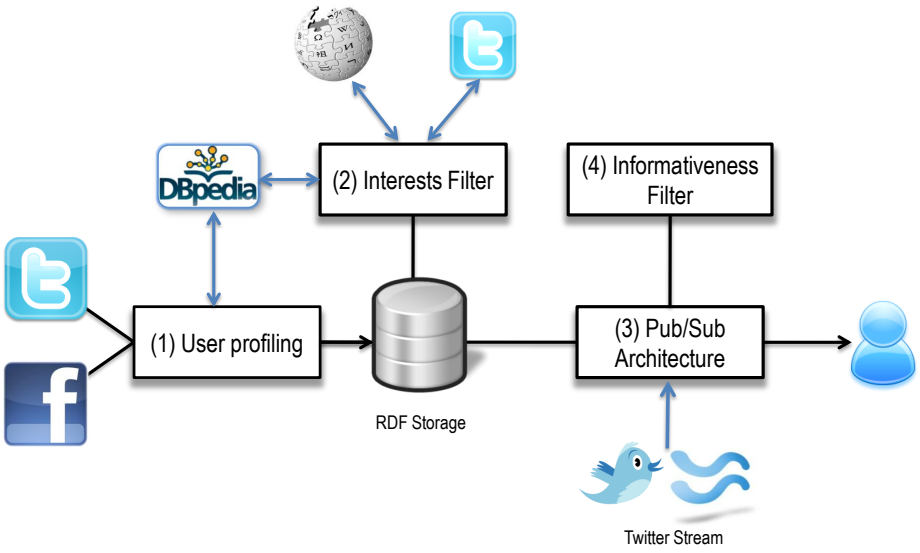

Figure 6.7.: Architecture

In order to build the system, we followed four steps (as illustrated in Figure 6.7):

1. generate multi-source provenance-aware user profiles of interests (Chapter 5);
2. select and categorise the most appropriate interests for filtering a large Social Web stream (Section 6.3.2.1);
3. analyse the stream and distribute the tweets to the interested users in real-time;
4. deliver to the users only the most informative tweets for each topic (Section 6.3.2.2).

Each step is represented by a specific software module as displayed in Figure 6.7. The Pub/Sub architecture is based on the Semantic Hub (SemHub — [Kapanipathi et al., 2011a]): an extension of Google's PubSubHubbub (PuSH — the Google's Publish/Subscribe protocol) using Semantic Web technologies to provide publisher-controlled real-time notifications. Further details about our work (in collaboration with Kapanipathi and Sheth) on this distributed and scalable software architecture have been published in [Kapanipathi et al., 2011b].

6.3.2.1. Filtering User Profiles of Interests for Personalisation

In order to provide real-time personalisation of a large social stream and avoid overload of information it is necessary to have a recommendation system able to select only the few very relevant entities of interest to provide recommendations for. After a preliminary

experiment with a filtering system for Twitter, we realised that filtering and enriching the concepts of interest are fundamental steps for reducing a large number of resources populating user profiles. Otherwise, the usage of unfiltered profiles would produce simply more overload of recommendations. We introduced some measures for characterising concepts of interest (Section 6.2) and here we describe a methodology for combining those measures and improve recommendations' quality in our use case. We propose to use *specificity* and *popularity* respectively as filters for deciding whether the interests are useful or not for personalisation and whether their related streams of messages need to be filtered by informativeness measures or not. Figure 6.8 clarifies this theory. Briefly, if an entity is identified as popular and generic then we discard it as it generates many tweets and it is also about something abstract so probably very noisy. On the other hand, if a concept is also popular but specific then we keep it, as it might be relevant. We then have to filter the related tweets using informativeness measures, as in Section 6.3.2.2. For all the other cases we keep the entities in the user profiles as they anyway do not generate a high number of posts.

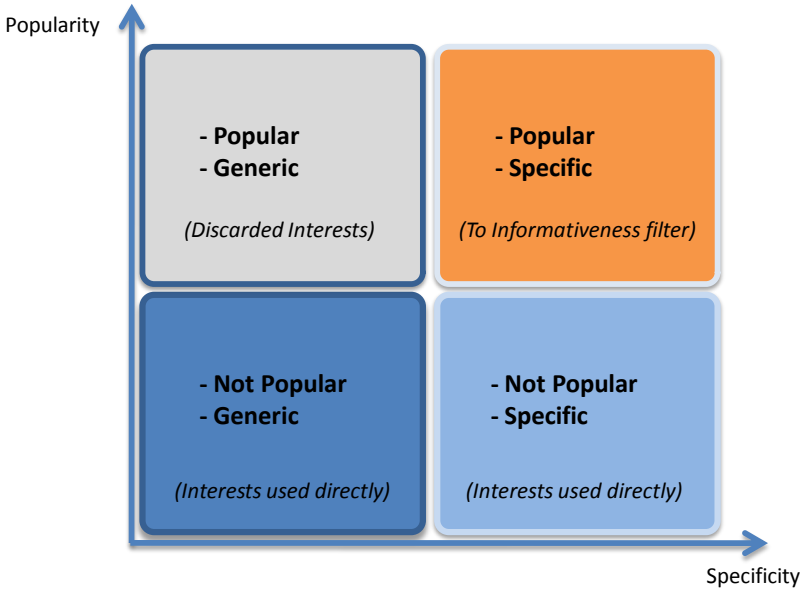

Figure 6.8.: Combining popularity and specificity for filtering the interests.

As regards the temporal dynamics of the interests we consider those two measures ("*stable*" and "*trendy*") as features useful for modifying the relevance weight of the interests in a user's profile. Hence, if an entity is "not stable" we increase its weight, and if an entity is "trendy" we also increase its weight. This is because we observed that those two categories of concepts receive more attention from the users than others and, logically, it is highly probable that those concepts are related to some events of interest

on the Social Web. The increase of weight is useful because after this filtering process only a selected number of top interests are being considered by the recommender system. An evaluation of the impact of these measures and the methodology has been performed with the SPOTS system for the Twitter use case and is described in Section 6.3.3.

6.3.2.2. Informativeness of Microposts

Information overload, especially on popular social media services such as Twitter, is an ongoing problem. The users and the number of microposts (tweets) generated are consistently increasing. In this section, we introduce a measure that leverages intuitive features of a tweet to objectively score its informativeness. The contributions of the features used in this system for the informativeness of a tweet have already been proved by [Tao et al., 2012]. Although, the authors introduce many features, we restrict the feature set to the most prominent ones also by considering the real-time nature of our application. We reuse the work of Tao et al. and apply it to our use case which deals with filtering un-informative tweets. However, the work in [Tao et al., 2012] can be improved, especially considering our scenario, and it is part of our future work.

Link: Links (URLs) play a prominent role in containing the maximum information conveyed by a 140 characters tweet. Although only 21% of the tweets contain links [Hong et al., 2011], approximately 60-80% of retweets include a link[8]. Therefore, we consider the presence of a link in the tweet to be informative. Further, these links from the tweets have already been exploited in [Dong et al., 2010] to improve the temporal prominence of Web search results.

$$inf_{link} = \begin{cases} 1 & \text{if link present} \\ 0 & \text{if link absent} \end{cases}$$

Hashtag: Hashtags are a representation of a topic on Twitter and are present in around 19% of the tweets [Tao et al., 2012]. They are used as indexes to search and keep track of the latest happenings of a topic on Twitter. Further, they are also adopted by other social networks such as Google+ and Facebook[9]. Since users explicitly tag tweets

[8] See for example: http://techcrunch.com/2010/09/14/twitter-seeing-90-million-tweets-per-day/ and: http://goo.gl/5zsmX (accessed January 2014)

[9] http://duluth.patch.com/groups/editors-picks/p/facebook-making-more-changes-adopting-hashtag (accessed January 2014)

with a topic, the presence of a hashtag in a tweet improves the informativeness of the tweet. However, some tweets are tagged with more than one hashtag and some contain only hashtags. Studies have shown that the maximum engagement (21% higher[10]) of tweets are with one or two hashtags. Therefore, we progressively reduce the impact of increasing number of hashtags in a single tweet.
Formally:

$$inf_{tag} = \sum_{i=1}^{M} \frac{1}{2^i}$$

where M is the total number of hashtags in the tweet.

This measure has been suggested by the work of [Tao et al., 2012], however it can be improved considerably. Using this formula, we would still increase the informativeness score for an increasing number of hashtags, even though after three or four hashtags the value of inf_{tag} does not increase much. In this case it would probably be better to use a Gaussian function centred on the number 2, so that it would give more weight to tweets which have one to three hashtags. In our future work we plan to improve Tao et al.'s work and experiment with different functions.

Named Entities: The importance of semantic features such as the presence of Named Entities in the Tweet to make the tweet interesting has already been hypothesized and proved in [Tao et al., 2012]. Therefore, we consider the presence of the entities and the number of entities present in the tweet as one of the features to determine the informativeness. The score is determined by using the same formula used for Hashtags. However, the rationale behind was to provide more prominence to the first few entities found in a tweet. Also, during our experiments containing 50,000 tweets, the maximum number of entities extracted for a single tweet was 5.
Formally:

$$inf_{entity} = \sum_{i=1}^{M} \frac{1}{2^i}$$

where M is the total number of entities extracted from the tweet.

The same considerations as for the case of the hashtags can be done here with named entities. Probably, this formula, and Tao et al.'s work, can be improved with a Gaussian function similar to the one mentioned earlier for the hashtags. This however requires additional extensive experiments for its evaluation and it is part of our future work.

[10] http://www.mediabistro.com/alltwitter/twitter-strategy_b24623 (accessed January 2014)

Tweet Length: The restriction of tweet length to 140 characters makes it challenging for users to compose tweets and to appropriately convey as much information as possible. According to [Jenders et al., 2013] the most retweeted tweets, hence the most interesting ones on Twitter, have an average length between 120 and 130 characters. Therefore, we use the normalized length of the tweet as one of the features to determine the informativeness of the tweet.
Formally the normalized length is:

$$inf_{len} = \frac{length(tweet)}{140}.$$

To be more accurate, following [Jenders et al., 2013], we could have used a function to give higher score to tweets around 120/130 characters, and discount the score for tweets reaching the length limit of 140 characters. However, this needs an extensive evaluation and for the time being we decided to keep our function simple.

Retweets and Replies: Since the objective of the system is to provide new, non-redundant and informative tweets, we discard tweets that are either retweets (assuming retweets are tweets that are already processed) or replies (conversations) to another tweet. The retweets and replies constitute around 29% of the tweets generated everyday[11] so with this feature we significantly reduce the processed stream.

Aggregated Informativeness: Finally, we sum up the scores from each feature to return a composite score for the informativeness of the tweets as shown in the following equation:

$$inf_{tweet} = (w_{link} * inf_{link}) + (w_{entity} * inf_{entity}) + (w_{tag} * inf_{tag}) + (w_{len} * inf_{len})$$

Where:

$$w_{link} + w_{entity} + w_{tag} + w_{len} = 1$$

and each variable w represents a weight to give more or less relevance to one factor of the entire formula. However in our work we experimented only with weights w all equal and set to 0.25. The weights for the formula could be tuned using for instance a test dataset of tweets evaluated by human judges. Once we know which tweets are informative for humans, we can use the judgements in a dataset to tune the weights of the informativeness formula. Alternatively, we can use a dataset of tweets classified

[11] http://mashable.com/2010/09/29/twitter-replies-retweets/ (accessed January 2014)

by humans as training set for a machine learning algorithm to infer the weights for the informativeness formula. This is currently part of our future work, as an improvement over the research described in [Tao et al., 2012].

6.3.3. Evaluation Against Twitter's Recommendations

In this section we present the results of the evaluation of our system. We evaluate the quality of the recommendations of SPOTS with different setup conditions and compare it to the Twitter's *Discover* service. Twitter Discover is a public and commercial service offered to Twitter users and (as of November 2013) it is the only free service providing tweets recommendations for the entire public Twitter stream. We have used the limited stream provided by Twitter which represents the 10% of the complete public Twitter stream. However, the Discover service does not provide real-time recommendations, the tweets recommended span from a few days, to some hours, to a few minutes old. Its implementation details are unavailable but it aims mostly at recommending tweets from reputable sources or popular accounts or users' social connections (*e.g.* friends of friends). Also there are no ways to collect Discover's recommended tweets, so we had to ask the users of our evaluation to provide access to their Discover page for collection. We collected tweets at most 10 minutes old, as old tweets give the advantage that the popularity and distribution of the tweets (e.g. if retweeted or favourited) could be used for computing their relevance.

For the evaluation setup, we focused on standard protocols for recommender systems [Herlocker et al., 2004]. We selected 7 active users from the user survey about the profiles of interests previously described in Chapter 5. We generated their user profiles from Twitter and Facebook and asked them to give us access to their Discover page. Just after the collection of tweets from Discover we started the SPOT system and we allowed the system to collect tweets in real-time for a period of 5 minutes in order to have enough tweets for the recommendations for every interest in the user profiles. SPOTS has been used with user profiles containing only the top 10 interests which were not filtered using the interests filter, so the ranking was based on occurrences, time decay and provenance-based features only (as described in Section 5.3). For each interest we selected and provided 3 tweets that were then given to the users for evaluation together with 10 Twitter Discover tweets in a randomised list. So in total every user evaluated 40 tweets (3*10 from SPOTS + 10 from Discover), therefore 210 tweets were evaluated

for SPOTS and 70 for Discover. Users were asked to provide a score from 1 (low) to 10 (high) for each tweet according to their interest.

In Table 6.2 we display the results of the evaluation. As we can see from the average scores obtained from the evaluation, SPOTS overall (even without any filter for the interests) provides better recommendations than the Twitter Discover service (average score 6.34 for SPOTS *versus* 3.29 for Discover). The other strategies displayed in the table represent different interest filtering techniques applied to SPOTS. We can see that if we use only *Trendy* or *Unstable* interests we have the best performance in the recommendations. However the amount of concepts categorised in one of these two classes is limited and users would loose many relevant interests not belonging to these classes. Hence, we decided to combine features for filtering in order to obtain a trade-off between accuracy and variety of interests. With "SPOTS w/ at least 2 features" we consider only interests from the user profiles which are belonging to at least 2 of the following categories: *Trendy*, *Unstable*, *Specific*, *Not Popular*. Statistical significance of these different results has been tested with a two-tailed unpaired t-test and $p < 0.05$. In the same way, with "Specific + Not Popular" SPOTS has been configured to use only *Specific* and *Not Popular* interests. The filter on "Trend" selects on average only 6% of the interests, the one for "Not Stable" 20%, "Specific + Not Popular" 36% and "at least 2 features" 38%.

The advantage of filtering user profiles of interests by enriching them with external information sources is clearly demonstrated by the results of this evaluation. Features for characterising and filtering user interests, such as specificity, popularity and its trend, are also useful for improving recommendations, as shown with this type of personalisation use case described in this section. A thorough and more generalised analysis of these features for semantic enrichment of user profiles of interests is provided in Section 6.4.

6.4. Evaluating Aggregated Provenance-Aware Semantic User Profiles

6.4.1. Semantic Enrichment User Study

In this chapter we have proposed a model and a set of measures for characterising entities of interest. In this section we present an evaluation of the impact of this model

System and Strategy	AVG Score	std.dev.
SPOTS without filtering	6.34	1.21
Twitter Discover	3.29	1.48
SPOTS only Trendy	8.89	1.17
SPOTS only Unstable	8.60	1.64
SPOTS w/ at least 2 features	7.13	1.23
SPOTS Specific + Not Popular	6.73	1.93
SPOTS only Specific	6.53	1.44
SPOTS only Not Popular	6.52	1.98

Table 6.2.: Average scores for the recommendation systems SPOTS and Twitter Discover and impact on the scores due to different interest filtering strategies (1 to 10 scale).

and the semantic enrichment directly on user profiles of entities of interest. We generated user profiles for 27 users (volunteers for our user study) as previously described in Section 5.4.2.2 and for the evaluation of SPOTS in Section 6.3.3. Each user was asked to rate the relevance of 30 entities of interest according to their personal preferences. The entities of interest were generated and ranked for their user profile according to their activities on Facebook and Twitter and their number of mentions in their social data (occurrence-based weighting strategy). Hence, the ranking was based on occurrences, time decay and provenance-based features only (as described in Section 5.3). The methodology for the profile generation is the same as the one described in Chapter 5 (see also [Orlandi et al., 2012]). This user study aims at evaluating, directly with user judgement, the impact of our features for semantic enrichment on the accuracy of our profiling methodology. In total we collected 794 user ratings (not 810 because some users evaluated less than 30 entities) on a scale from 1 (low relevance) to 5 (high relevance), on a total of 529 distinct DBpedia resources as interests. For every entity we computed our measures for semantic characterisation as described in Section 6.2 and we analysed the average user score, grouping by each different feature.

As we can see from Table 6.3 the entities of interest categorised as "Non-Specific" (which have high values for our Specificity measure) provide an improvement on the user score of almost 8% on the average score for all the interests. This means that users perceive abstract concepts of interest as more relevant for their user profiles. This improvement has been confirmed by the tests for statistical significance performed (two-tailed unpaired t-test with $p < 0.05$). The other measures (Popularity, Temporal Stability

Type of entity	Tot. Entities	AVG Score	Std.dev.
All	794	3.34	1.47
Non-Specific	297	3.66	1.39
Non-Popular	410	3.40	1.46
Stable	663	3.37	1.47
Non-Trendy	778	3.35	1.47
Stable & Non-Trendy	659	3.38	1.47
Non-Popular & Non-Specific	134	**3.84**	1.39

Table 6.3.: Evaluation of the average user scores (on a 1 to 5 scale) grouped by type of entity of interest.

and Trend) do not show significant improvement on the average score. We also show the effect of aggregation of two types of interests: "Stable plus non-Trendy ones" and "non-Popular plus Non-Specific" concepts. The latter gives best results with more than 12% improvement over the average user score and demonstrates the validity and complementarity of these two measures. To note that the threshold chosen for the binary classification of every measure is the median of all the values for the measure. This, following some early empirical experiments, was the threshold maximising the accuracy of the classification.

With this user study we provided insight on the effect of these dimensions for semantic enrichment on user profiles of interests.

6.4.2. Overall Evaluation of Semantic Enrichment for Personalisation

While in Section 6.3.3 we provided an evaluation of characterisation and enrichment features applied to a personalisation use case such as real-time tweets recommendations, in the previous Section 6.4.1 we evaluated the same features through direct user feedback on their personalised user profiles. By comparing the two different evaluation scenarios, we observe a clear difference between the relevance that a user assigns to a concept of interest or to a personalised recommended object. From the results obtained with the two evaluations, users prefer on the one hand, to be categorised with abstract concepts of interest and, on the other hand, to receive recommendations related to specific interests. Moreover, measures of popularity and temporal dynamics of interests demonstrated to be more relevant in ranking the importance of the interests for Social Web recommendations than for user profile representation. In contrast, specificity measures are more useful for a representation of people's interests than for real-time Social Web

recommendations. In Table 6.4 we summarise these results for the two different types of evaluations. We illustrate an aggregation and comparison of the most important and statistically significant results obtained, as originally reported in Table 6.2 and 6.3. From the results of the semantic enrichment evaluation it is clear that a correct use of the proposed characterisation features could bring more than 25% improvement to the accuracy of recommendations and user models.

SPOTS	**Improvement**	**User Study**	**Improvement**
Trendy	+29%	Not Specific + Not Popular	+13%
Unstable	+26%	Not Specific	+8%
At Least 2 Features	+9%	Not Popular	+2%
Specific + Not Popular	+5%	Stable & Not Trendy	+1%

Table 6.4.: Average score improvement of semantic enrichment over non-enriched user profiles of interests for the two different evaluations: the recommender system SPOTS, and the user study.

6.5. Conclusions

In this chapter we described a methodology for semantic enrichment and deployment of user profiles of interest. This approach represents the last steps of our complete methodology for profiling user interests and it is built on top of the architecture for mining and aggregating user interests described in the previous chapter. We leverage Web of Data and Social Web for enriching the knowledge related to the entities in the user profiles. In particular, first we proposed a real-time, computationally inexpensive, domain independent model for characterising concepts of interest: based on specificity, popularity and temporal dynamics. Then, as the last step of our profiling methodology, we focused on the deployment of enriched user profiles on practical personalisation use cases. We evaluated our complete profiling methodology on a personalisation system (called SPOTS) implemented for real-time filtering of Social Web streams of messages (such as the Twitter stream). Specificity in particular revealed to be extremely important for user modelling and representation of interests on the Social Web, as we showed that user interests can be ranked also according to their conceptual level of abstraction. Trend and popularity of concepts on the Social Web can be considered complementary to specificity and provide insight on the semantics and pragmatics of the entities. Finally, this chapter, demonstrated how characterisation and filtering of the interests are strongly

dependent on the personalisation use case. Moreover, the proposed model of enrichment should adapt to every type of deployment of the profiles.

Chapter 7

Conclusions and Future Work

In this thesis we formalise a methodology for profiling user interests which leverages the Social and Semantic Web. Following an introduction to the problem of current Social Web personalisation systems, we described the proposed solution, and evaluated its deployment, following the main stages of a user profiling pipeline (as depicted in Figure 1.2). In this chapter we conclude the thesis recalling the research questions identified in Chapter 1 and discuss the results we have delivered (Section 7.1.1) as well as the important lessons we have learned when attempting to find the answers (Section 7.2). Our investigation on a new methodology for user profiling unveiled novel scenarios and additional research questions. In Section 7.2 we describe the planned continuation of our work, possible new goals and future developments derived from the work presented in the thesis.

7.1. Conclusions

The core research question of the thesis, introduced in Chapter 1, expresses in a generic way the main goal of this thesis: investigating, formalising and evaluating a methodology for profiling user interests on the Social Semantic Web. The main research question is as follows:

> *How can we collect, represent, aggregate, mine, enrich and deploy user profiles of interests on the Social Web for multi-source personalisation?*

We divided the main methodology and our investigation into three parts, each one identified by a more specific research question (as described in Section 1.2 and in the following Section 7.1.1). As illustrated in Figure 1.2, the first research question is connected to the first four stages of the proposed profiling methodology: collection, representation, aggregation and mining. This question (Chapters 3 and 5) investigates the basis of the profiling pipeline which aims at generating sets of relevant interests for the users, extracted from multiple social media sources. The second question is about provenance of data, which involves all the stages of the profiling methodology. In Chapter 4 we demonstrate the importance of provenance of data for user profiling and personalisation. The third and last question (Chapter 6) investigates the semantic enrichment of the user profiles and their deployment for personalisation use cases.

We summarise the answers to these questions in Section 7.1.1. The main outcome is a complete methodology for profiling user interests that goes from the collection and aggregation of user data from heterogeneous Social Web platforms, to the management and representation of this data, to the semantic enrichment of interoperable user profiles ready to be adapted and deployed for different personalisation tasks.

7.1.1. Answering the Research Questions

1. **Aggregation of Social Web data for profiling user interests**: *How can we aggregate and represent user data distributed across heterogeneous social media systems for profiling user interests?*

   The importance of the aggregation of Social Web data for mining user interests has been emphasized in Chapter 3. In the same chapter, we describe the main challenges for aggregation of heterogeneous social networking systems and user modelling on the Social Web. These challenges reside on the lack of interoperability among Social Web systems and the high diversity of social media systems and user activities. Therefore with this thesis, we first propose a characterisation of the main types of social media systems and the different Social Web activities that users can perform on them. We base our characterisation on existing popular vocabularies for representing Social Web content in a structured format (e.g. SIOC, FOAF, Activity Streams, etc.). We describe the advantage of using popular standard Semantic Web ontologies for representing Social Web data and solving interoperability issues (Chapter 3). Subsequently, we propose a methodology based on the extraction of complete Social Web data and provenance information about users' Web activities

directly from the heterogeneous Social Web services available. We detail an efficient semantic modelling solution for Social Web data, based on the SIOC ontology and its extensions, and demonstrate the potential of a unified and interoperable meta-level description of users' Social Web data. On top of this structured semantic layer it is possible to perform an analysis of the different kinds of Social Web activities and user generated content in order to retrieve concepts of interest.

We implemented our methodology, for evaluation purposes, in a particular use case aiming at interlinking heterogeneous online wiki systems (Chapter 3). The experiment shows the applicability of our approach not only to wikis but also to other social media sites. In fact, we extend the implementation — and evaluation — of our methodology for user profiling also to other social media such as microblogs and social networking sites (as described in Chapter 5). The investigation conducted for this research question unfolded the other two questions. One related to the management and utilisation of provenance of data for mining user interests, and the other one related to the semantic enrichment and deployment of user profiles of interests for personalisation.

2. **Provenance of data for user profiling**: *What is the role of provenance on the Social Web and on the Web of Data and how to leverage its potential for user profiling?*

   Provenance of data plays a crucial role in social media and the Web of Data. In Chapter 4 we showed how provenance of data can be recorded and represented on the Social Web, and consequently used on Linked Data to track the origins of particular statements and resources. Similarly, provenance on/for the Web of Data can be used in many different use cases supporting Social Web users. For example for enriching user profiling processes or for computing trust and data quality measures on the Social Web.

   As an experiment, in Chapter 4, we focus on a particular use case involving Wikipedia and DBpedia. We propose a methodology to semantically represent information about provenance of data in DBpedia and an extraction framework capable of computing provenance for DBpedia statements using Wikipedia edits. Then, by indicating by whom and when a statement was created, we let any Social Web application evaluate DBpedia statements based on particular criteria. This example is a demonstration of interconnection and mutual dependence between Social Web and Web of Data, which we extend with a generic modelling solution for

provenance on the Social Semantic Web. This solution is compliant with the W3C standard PROV vocabulary for the representation of provenance on the Web and nicely integrated with popular Social Web ontologies such as SIOC and FOAF.

In the thesis, we identify provenance of data as the "glue" that could connect the Social Web and the Web of Data. Additionally, we demonstrate the potential of semantic representation of provenance in profiling user interests. One of our goals is to analyse the impact of provenance information and different types of Social Web features on automatically generated user profiles of interests. In this regard, at the end of Chapter 4, we identify potential provenance features that could be used for enhancing our user profiling methodology. Afterwards, in Chapter 5, we describe how we model provenance information extracted from different social media platforms. We detail how we integrate this provenance information in a user profiling pipeline and we evaluate different heuristics for mining user interests using provenance and Social Web data. In Chapter 6 we evaluate additional semantic enrichment techniques for concepts of interest and also evaluate the impact of the aforementioned provenance features on personalisation scenarios.

3. **Semantic enrichment of user profiles and personalisation**: *How can we combine data from the Social and Semantic Web for enriching user profiles of interests and deploying them to different personalisation tasks?*

   Following the analysis on how to mine, aggregate and represent user interests from different sources on the Social Web, we focused on enrichment and deployment of user profiles of interests. In Chapter 5, semantic representation of concepts of interest and the Web of Data demonstrated to be crucial for selecting, ranking and filtering user interests according to several measures and the particular use case. By exploring the Linked Data graph, we can relate information to the original concepts of interest of any of our entity-based profiles and enrich their semantics. In Chapter 6 we describe our methodology for semantic enrichment and characterisation of concepts of interest. We employ the Web of Data and the Social Web for the enrichment and evaluate the impact of our measures on selected personalisation scenarios.

   As regards the enrichment of the profiles, we use the Web of Data not only to link to its entities but also to extract related resources and categories for concept expansion. We then leverage the structure of the Linked Data graph, together with the vast and timely knowledge on the Social Web, to better understand the appro-

priateness of some of its entities for representing user interests. Hence, we propose a novel approach using Social Web and Linked Data information for characterising important dimensions of entities of interest: specificity, popularity and temporal dynamics. From our user studies (Chapter 6) these features resulted relevant in general for Social Web users and also in particular for a specific personalisation use case. Specificity is extremely relevant for use modelling and representation of interests on the Social Web as we showed that user interests can be ranked also according to their conceptual level of abstraction. Trend and popularity of concepts on the Social Web can be considered complementary to specificity and provide insight on the semantics and pragmatics of the entities.

These features of concept characterisation proved to be essential, for instance, in filtering and ranking preferences in real-time over a large Social Web stream of messages. In fact, this particular scenario has been adopted for evaluating how to deploy user profiles on practical personalisation use cases. We propose a methodology and a set of heuristics to filter any public and large social stream of short messages and personalise it, in real-time, according to automatically updated user profiles of interests. We describe the theoretical background and the implementation of "SPOTS" a system offering real-time personalisation of the public Twitter stream. SPOTS aims at recommending interesting tweets to users according to semantically enriched user profiles and specific informativeness measures. We provide a user-centric evaluation of our personalisation system SPOTS and a study on the impact of our profiling methodology on a real-time personalisation system.

Our methodology for user interests profiling has been evaluated in two different ways through user studies. One way aimed at investigating the impact of semantic enrichment on the accuracy of user profiles by asking for feedback directly from the users. The other way evaluated the same impact on the accuracy of user profiles in a personalisation use case. Hence, we analysed the user ratings given to the recommendations provided by our system SPOTS, which was fed with the same user profiles of the other user study. The results suggest that abstract entities provide better scores when user profiles are evaluated by the users themselves and that specificity and popularity positively complement each other. The same study applied to real-time tweets recommendations demonstrated the importance of semantic enrichment and interest filtering as essential phases of a user profiling process. The interests characterisation and filtering phase is strongly dependent on

the personalisation use case and the proposed features can easily adapt to every practical scenario.

7.2. Lessons Learned and Future Work

In this section we discuss some of the critical points encountered in our research, the related lessons learned and possible future work. As our investigation focused on the three research questions, we keep the structure given by the questions and divide our discussion accordingly.

1. **Federated Personal Data Manager**

   One of the main obstacles to the aggregation of distributed Social Web data is the lack of interoperability of current social media. Most of the Social Web services clearly aim at keeping their users and their data closed inside their platform. This is due to their business model: the main source of income for many social platforms is depending on the number of their users and how many times their advertisement is being "watched". For this reason our methodology needs to collect user data from the Social Web sources and represent it using an interoperable and standard format. This additional abstraction layer is necessary for interoperability on the Social Web and other Web applications can use it and interact with it.

   The critical point here is on the possible applications that would benefit from such user profiles. We believe that the need for a privacy-aware personal profile manager for Social Web users will increase in the near future. This kind of application would support users in managing their social data efficiently and specifying privacy preferences for applications or users requesting access to the users' data. In Section 5.5 we have introduced this problem and a similar application, together with related work that shares our vision. This kind of application, or manager for user profiles of interests, is important for the consequent development of other third-party applications that would benefit from exchanging data with this personal manager. For example, Web applications that need accurate user data for providing their personalisation services could get complete multi-source user profiles from the users' managers and, at the same time, contribute in enriching the profiles with their own user information. The whole process would take place in complete agreement and control of the user who would just need to specify her privacy preferences. In order

to pursue this goal in the future, it will be necessary to give the social media service providers sufficient motivation for "opening up" their data to their users so that users and service providers could collaborate for fair and accurate personalisation.

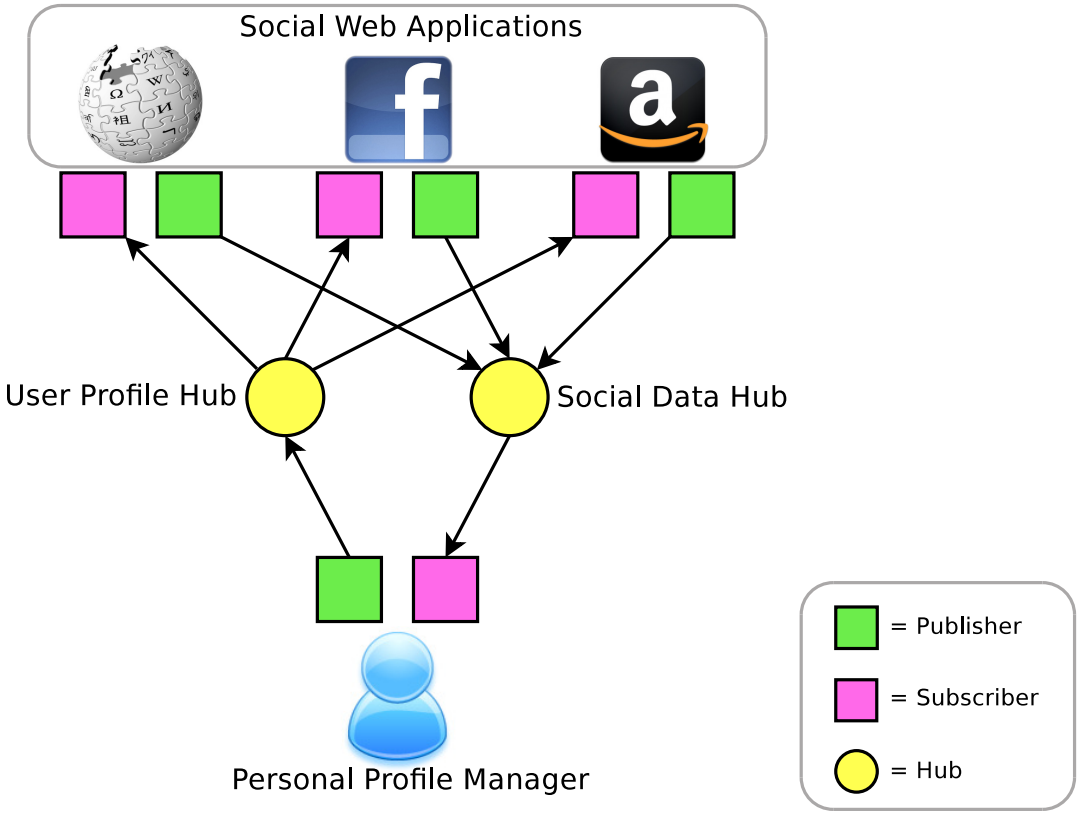

Figure 7.1.: Example of a possible federated architecture for user profile distribution and personalisation. It is based on a pub-sub protocol where the User Profile Hub receives updates on the user profile generated by the Personal Profile Manager and distributes them to Social Web applications for personalisation. At the same time, the Social Data Hub receives updates on the Social Web activity of the user and distributes it to the Personal Profile Manager of the user for profiling.

In our future work, we will investigate the potential of such an interoperable personal manager for user profiles of interests. In particular we envision a federated lightweight semantic architecture for the manager and the applications for personalisation. This would be possible and efficient with a pub-sub architecture. An example is depicted in Figure 7.1. Web applications could subscribe to the User Profile Hub of a user in order to get updated user profile information. Vice-versa, a user's Personal Profile Manager could subscribe to her Social Web applications for updates on her social media feeds. These updates would then be used to dynamically refine the user profile. The pub-sub architecture guarantees real-time capabilities, decentralisation and flexibility.

2. **Provenance at Web Scale**

As described in Chapter 4, provenance of data is essential for tracking the origin of Social Web data. However, the collection of complete provenance of data is often a non-trivial task. In Section 2.2 and 4.2 we describe the main challenges of provenance of Social Web data, in particular:

- *Provenance on social media is hard to track.* Currently, on the Social Web, an efficient and standard approach for reconstructing information propagation, tracking the possible diffusion of information on social media, is not available. In this thesis we do not tackle this challenge and we refer to some early related work in the field[1]. We assume that the provenance information that we can collect from social media (in the same way as we described for the Wikipedia, Facebook and Twitter cases) is always correct. In fact, the knowledge about how a piece of information was modified and propagated through social media and how an owner is connected to its transmission is often missing. Therefore there is a need for accurate and complete provenance information available and extended to Web scale. For instance, social media providers, who make APIs available, should provide also provenance metadata. This aspect is essential for user profiling and it is part of our future work.

- *Provenance on the Semantic Web can be expensive.* As estimated with our experiment, described in Section 4.3.1, a complete provenance information dataset for the English Wikipedia — expressed in RDF and using our modelling solution — would consist of 7 billion RDF statements. This is an extremely big number of triples to manage with the current storage technologies and it exceeds the number of triples dedicated only to the Wikipedia itself without provenance information. A better modelling solution for provenance on the Semantic Web has been investigated by the W3C Provenance Working Group[2] and the Semantic Web community. For a large number of RDF statements, solutions such as N-Quads[3] or Named Graph[4] could help in reducing this number. For instance, provenance statements can be referred to an entire group of

[1] Especially in [Barbier et al., 2013] and the work of the W3C Provenance Working Group [Gil et al., 2010]

[2] The W3C Provenance Working Group closed its activity on the 19th of June, 2013: http://www.w3.org/2011/prov/wiki/Main_Page (accessed January 2014)

[3] RDF 1.1 N-Quads is a line-based syntax for an RDF datasets. It is a W3C Proposed Recommendation published on 09 January 2014: http://www.w3.org/TR/n-quads/

[4] Named Graphs have been included in the recent RDF 1.1 Recommendation: http://www.w3.org/TR/rdf11-concepts/#dfn-named-graph. The W3C Prov-WG published important requirements for

statements enclosed in a named graph, and/or the additional fourth parameter of a "quad" could be used for provenance purposes.

3. **Adaptive Profiling of User Interests**

The part of the thesis that involves semantic enrichment of user profiles and their deployment for personalisation use cases is the one that opens a larger research area and more future directions. In particular the scientific investigation on the Linked Data graph for personalisation shows great potential and it is still largely unexplored. Promising research has been recently focusing on spreading activation techniques over the Linked Data cloud in order to navigate its graph and provide novel interesting recommendations [Marie et al., 2013] [Heitmann et al., 2012c]. Not only the academic world but also industry and popular companies such as Google and Facebook demonstrated their interest in this field. In this thesis we have shown how to expand the knowledge about user interests by linking to the vast and open Linked Data cloud. By analysing its graph structure and the semantics of its concepts it is possible to extract useful measures for identifying the most appropriate resources to use in different personalisation contexts (Chapter 6).

One of the lessons learned with our experiments on the semantic enrichment is about the "noisy" nature of the Web of Data. Dealing with such a vast, dynamic and open corpus such as the LOD cloud is a non-trivial task and often leads to the addition of too much noisy or unnecessary information to the experiment dataset. For this reason we proposed measures for the Linked Data graph that are simple and suitable for real-time computing. Further studies will be conducted on the development of novel measures for the characterisation of entities of interest. These features will have to be tested on different personalisation scenarios as their impact can change depending on the use case, as shown in this thesis.

As regards the deployment of the user profiles for personalisation, in the thesis, we support the development of methodologies for filtering user interests in a profile according to the personalisation task and user context. Future studies will focus on the development of strategies for the adaptation of the complete profiling algorithm according to the personalisation use case. Being it for movies recommendations or for filtering blog posts, the deployment module should be adaptive and designed to select only the user interests that are relevant for the specific task. This research

Named Graphs here: http://www.w3.org/2011/prov/wiki/ProvenanceRDFNamedGraph (accessed January 2014)

challenge meets our other idea, previously described in this section, about a personal profile manager. For example, a personal user profiling agent could provide an automated and dynamic profiling service to the user and, at the same time, it would offer filtered and personalised user profiles to the applications asking for user information. The profile provided to an application would be tailored to the specific required purpose.

An additional future direction for our work would be the adaptation of the profile according to the user context, not only to the personalisation use case. A user profile expressing user interests should adapt also to the location, environment, current activity and time of the user who needs it for real-time personalisation. Our study could be extended also to the use of smartphones which would help in detecting the context of the user thanks to their built-in sensors. As a simple example, we could imagine the user profiling algorithm adapting the interests when the user is at home relaxing or at work. This would enable the creation of a semantic, autonomous, software agent for the management of personal Social Web information.

Part I

Appendix

Appendix A

User Studies on the User Profile and Privacy Manager

A.1. Preliminary User Study on Privacy for User Profiles

This user study has been published in collaboration with Owen Sacco in [Sacco et al., 2012]. Prior to implementing the privacy preference manager that provides users to create privacy preferences for generating faceted profiles, we first conducted an online survey in order to understand what users think about protecting their personal information published online. This survey serves as the requirements for designing our interface; to know which options to provide to end-users. The survey contains 7 questions which, together with the results from 70 users, are illustrated in Figures A.1 - A.7.

Question 1 (Figure A.1) shows that 98.60% of the users are aware of privacy settings since they have set them at least once in current Social Web applications. The user who said no and the other user who skipped this question informed us that they are not confident in publishing information in current Social Web applications due to privacy issues, and hence, they do not use these type of applications. This illustrates that users

1. Have you set at least once your privacy settings on your Social Web application of your choice (such as Facebook, LinkedIn or Google+,etc..)?	Yes	No
	98.60%	1.40%

Figure A.1.: Privacy Preferences User Study - Question 1

2. Do you share your profile information (such as interests, contact information, demographic information etc) to everyone or to a restricted number of users?	Everyone	Restricted number of Users
	11.40%	88.60%

Figure A.2.: Privacy Preferences User Study - Question 2

3. If provided by the system, would you set different privacy settings for each part of your profile information? For example: a privacy setting to grant access to your family members to see your personal mobile number and another privacy setting to grant access to your work colleagues to see your email address.	Yes	No
	92.90%	7.10%

Figure A.3.: Privacy Preferences User Study - Question 3

are unhappy with current implementations of privacy settings. Question 2 (Figure A.2) illustrates 88.60% of the users are unhappy to share their profile data with everyone and prefer to grant access to a restricted number of users. Therefore, this shows that users require to set privacy settings for their profile information. Question 3 (Figure A.3) demonstrates that 92.90% require to have fine-grained privacy settings for their personal information which current Social Web applications do not provide.

In question 4 (Figure A.4) we asked the users to which parts of their profile information they will most likely set fine-grained preferences. All the attributes contained within the list were chosen revealing that users require to set fine-grained privacy preference for each single information contained in their profile; contact information such as phone numbers and also photos being the most required by 97-95% of the users. 5% of the users provided us with feedback mentioning that they would set different privacy preferences for status messages and micro-posts since they feel confident with publishing micro-posts to a larger audience and they are more concerned to whom they share their status messages. This illustrates that users require fine-grained privacy preferences for their status messages. Question 5 (Figure A.5) demonstrates that 66.70% are willing to set fine grained privacy settings more than once which shows the importance of having a scalable system that provides users to set restrictions to whom they share information with.

In question 6, we asked which attributes users requesting personal information must have in order to share with them private sensitive information. 82.30% of the users

4. If provided by the system, to which attributes will you set fine-grained privacy preferences?	
Nickname	22.10%
Full Name	33.80%
Gender	22.10%
Birthdate	63.20%
Email	85.30%
Mobile / Phone Number	97.10%
Photos	95.60%
Publications	35.30%
Homepage	27.90%
Contact List	76.50%
Location	64.70%
Interests	45.60%
Online Accounts	76.50%
Education	33.80%
Affiliations	36.80%
Projects	44.10%
Status Messages / Micro-posts	73.50%

Figure A.4.: Privacy Preferences User Study - Question 4

5. If the system provides fine-grained privacy settings for each part of your user profile information, how often would you set your settings?	
Never	1.40%
Only Once	21.70%
Occasionally	66.70%
Frequently	10.10%

Figure A.5.: Privacy Preferences User Study - Question 5

6. If provided by the system, would you grant access to other users based on the following attributes:	
Nickname	21.00%
Full Name	48.40%
Age	21.00%
Email	38.70%
Homepage	22.60%
Users in your contact list	82.30%
Location	35.50%
Interests	37.10%
Online Accounts	29.00%
Education	29.00%
Affiliations	48.40%

Figure A.6.: Privacy Preferences User Study - Question 6

7. If provided by the system, would you grant access to parts of your profile information to users who you don't know but have similar attributes (for instance interests) as yourself?	Yes	No
	43.50%	56.50%

Figure A.7.: Privacy Preferences User Study - Question 7

answered that they feel confident with sharing information to users in their contact list. Our hypothesis to this result is that users are used to this option since current Social Web applications provide to restrict their information based on contact lists. In order to verify our hypothesis, we omitted to have a contact list in our system but provide users to specify to whom they share information based on similar attributes to theirs. Question 7 inquired whether users prefer to share personal information with users who they don't know but based on similar attributes to theirs, or to users who they already know. Although the results revealed that 56.50% feel more confident in sharing information with people who they know, 43.50% reveal that people are willing to share their information based on similar attributes to people who they don't know. Since the results are almost equal, this also encourages us to develop a system without any contact lists.

A.2. Evaluation of the System for User Profile and Privacy Management

The evaluation of our system involved users to create privacy preferences and verifying that what they created corresponds to what other users are allowed to view. The process of the evaluation consisted of a one-to-one interview whereby we commenced by explaining our objectives and overview of our work. We then asked the users to perform 3 tasks which consisted of the following:

(1) Create 2 or more attributes to users who work at the same workplace as yours;

(2) Create 2 or more attributes to users who are interested in a particular topic; and

(3) Verify how other users view part of your profile based on your privacy preferences.

After the users had completed these tasks, they were asked to complete an online survey which, together with the results, are illustrated in Figures A.8 - A.12. The users did not have any problems in getting used to the system. In fact, it took the users between 1 - 2 minutes to complete all the tasks. However, the interviews lasted between 20 to 45 minutes because in each interview each user provided feedback and was eager to try more privacy preferences than the amount specified in the tasks. Currently only 7 users were interviewed but we plan to extend the evaluation to include more participants.

Question 1 (Figure A.8) asked whether the system provided enough properties to conduct the task of creating privacy preferences and viewing faceted profiles. 85.70% of the users were satisfied with the options, however, 14.30% of the users stated that some of the interests were irrelevant and preferred to have an option to add/delete interests. Moreover, they also stated that they would have also preferred to have options to add specific users or user groups. In question 2 (Figure A.8), 71.40% state that the user interface was user-friendly, however, 28.60% of the users found that the interface provided long lists of interests which required the user having to select many interests. They suggested that interests should be grouped and categorised so that when a category is selected, all the interests in that category are also selected to be shared. Moreover, a user preferred that first they would like to select to whom they want to share first rather than first selecting what they want to share. This requirement is useful to improve the interface by catering for personalisation of user interfaces whereby each user can customise the interface according to their personal preferences.

1. Did the system provide you with necessary options to set your privacy preferences?	Yes	No
	85.70%	14.30%

Figure A.8.: MyPrivacyManager, User Evaluation - Question 1

2. Did you find the user interface easy-to-use to define fine-grained preferences?	Yes	No
	71.40%	28.60%

Figure A.9.: MyPrivacyManager, User Evaluation - Question 2

Question 3 (Figure A.10) shows that 57.10% of the users require more attributes to share such as photos. This means that the users are eager to use this system to create privacy preferences for more information and not only the ones collected from Twitter, Facebook or LinkedIn. Question 4 (Figure A.11) demonstrates that 42.90% of the users required more attributes such as location to specify to whom they want to share information. Most of the users suggested to retrieve more interests and not only the ones which they were interested in. Additionally, 42.90% of the users were satisfied with the attributes the system provided.

Question 5 (Figure A.11) illustrates that all users who were interviewed were satisfied with how the system filtered their profile and how the system generated the faceted profiles for different requesters. This verifies that the system generates the right faceted profile as how the user expected whilst creating their privacy preference.

Question 6 (Figure A.13) inquired whether the users would use the concept of creating and managing fined-grained privacy preferences for all their personal information on the Social Web. 85.70% answered that they were in favour of creating such fine-grained privacy preferences. This result encourages us to enhance and improve our system to provide as many options as possible for users to be able to create privacy preferences for any data collected and structured from the Social Web. 14.30% would not use this

3. Do you require more or less attributes to share?	
More	57.10%
Less	14.30%
Fine	28.60%

Figure A.10.: MyPrivacyManager, User Evaluation - Question 3

4. Do you require more or less attributes to specify the users to whom you will grant access?	
More	42.90%
Less	14.30%
Fine	42.90%

Figure A.11.: MyPrivacyManager, User Evaluation - Question 4

5. Did the preview of your faceted profile showed the correct information as how you expected?	Yes	No
	100.00%	0.00%

Figure A.12.: MyPrivacyManager, User Evaluation - Question 5

concept due to the tedious task of specifying many privacy preferences for each part of all their information published on the Social Web.

6. Once the system is improved and the user interface is enhanced, would you use this system to manage your privacy preferences for all your personal information on Social Web applications?	Yes	No
	85.70%	14.30%

Figure A.13.: MyPrivacyManager, User Evaluation - Question 6

Appendix B

Experiments for the Evaluation of the Specificity Measure

To evaluate our approach for identifying in real-time the specificity of entities on the Web of Data we tested the measures described in Section 6.2.2 on a set of 160 entities. The entities for our experiment were randomly selected from a large dataset of user profiles of interests generated for more than 50 different users. The profiles were automatically generated from the analysis of Facebook and Twitter user accounts as described in Chapter 5. For each entity we computed and recorded the value of different measures (our DRR, and the non-distinct similar ones: IP/OP, IP+OP, IP) by querying the Sindice SPARQL endpoint. We then compared those values with a gold standard generated by users classifying/rating the specificity levels of our test set of concepts. Additionally, we also reproduced a state of the art approach for measuring specificity based on the DMOZ hierarchical classification of the entities. We evaluate this other method against the gold standard and we then compare the accuracy of this method with our measure. The generation of the gold standard is described in the following section. The implementation of the DMOZ based method is detailed in Section B.2 and later in Section B.3 the evaluation and the results are examined.

B.1. Generation of the Gold Standard

Our gold standard has been generated through user manual annotation. The user evaluation set-up is composed of two interviews conducted at two different stages. **First**,

we asked 5 evaluators (2 females and 3 males, different age groups and expertise) to **classify** each of the 160 entities in two categories: *Specific* or *Generic* entities. As suggested to the evaluators: *"the classification should indicate whether the entity is an abstract concept in the real world and can be further refined and specified into many other levels of detail (hence Generic) or if it corresponds to a well defined and narrow instance (Specific)"*.

At a **second** stage (2 weeks later) we asked the same users to give **a score** to the same entities according to their perceived level of specificity. Instead of a binary classification then we looked for a more fine grained value. The scale used for the scores goes **from 1 to 10** (only integer numbers) where 1 identifies very generic entities (or with a very low level of specificity) and 10 was given to very specific entities.

The **first round of evaluation** has been completed by the evaluators on average in 20 minutes, while the second type of evaluation took more time: around 30 minutes. The second stage of the user evaluation has been conducted after the feedback received from the evaluators at the first stage and after a preliminary analysis of the results. Briefly, according to the users, in several cases it was difficult to choose between only two levels of abstraction, as entities have different degrees of specificity. The results of the 5 evaluators for the first evaluation have been aggregated and the inter-rater agreement has been computed. We computed the Fleiss' generalised Kappa coefficient for 160 subjects, 5 raters and 2 categories and we obtained $K = 0.61$. This value, according for example to the scale for Kappa's significance by Rietveld and van Hout (1993) is considered as indicator of substantial agreement [Eugenio, 2000]. The 5 raters agreement for this classification process could then be used as a gold standard for the first evaluation.

At this stage the five evaluators classified 38% *Generic* concepts and 62% *Specific*. As the entities collected for the evaluation are randomly extracted concepts from user profiles of interests, it is reasonable to have such percentages. Ideally, if we think about taxonomies of concepts the number of those which are generic, and hence on top of a hierarchical classification system, are less than the specific ones which are closer to the leaves of the hierarchical tree.

For the **second evaluation**, as previously introduced in this section, the same five users were asked to rate the specificity of the same 160 entities on a 1 to 10 scale of integers. For this type of evaluation it was not appropriate to compute the Kappa coefficient for the inter-raters agreement, as the number of categories in this case was

high (i.e. 10 categories). Hence, mean values and average standard deviation for the different ratings provided by the users were computed to estimate the agreement of the evaluators. In Table B.1 we provide details about this part of the evaluation. An analysis of the results will be provided in Section B.3.

Average Rate	7.03
Average Std. Dev.	1.45
Average Top30 High Std. Dev.	5.66
Average Top30 Low Std. Dev.	7.51

Table B.1.: Second evaluation, rating: Details about the scores given by the five users.

Interesting to note that the average standard deviation of the ratings is 1.45 on a scale of 10 values, which is an acceptable value. Moreover, the average score given by the raters is 7.03, which confirms again the tendency highlighted by the first evaluation of having a higher percentage of specific concepts. Interestingly, the average score for the top 30 entities with highest, or lowest, standard deviation is respectively 5.66 and 7.51. This clearly means that entities with the highest disagreement among the evaluators have lower scores and hence are more generic. A behaviour observed also in the first evaluation.

The purpose of this second different experiment is, first of all, to analyse the raters agreement in two different tasks, as suggested also by the results of the first experiment. Moreover, with the fine-grained ratings provided by the users we could rank the specificity of the entities, use this ranking as our gold standard and compare it to the other different ranking strategies given by our Linked Data measures and the DMOZ categorisation. As previously described, especially in a use-case scenario where user profiles of interests need to be ranked and filtered for selecting the top most relevant and specific interests, it is beneficial to have fine-grained values allowing for specificity ranking methods. More details about the results are described in Section B.3.

To evaluate the accuracy of the different ranking strategies we use the following prominent Information Retrieval ranking evaluation metric: the *Normalized Discounted Cumulative Gain (NDCG)* [Järvelin and Kekäläinen, 2000]. This evaluation metric supports graded judgments and penalizes error near the beginning of most relevant tags determined by our approach. NDCG is the normalized value of Discounted Cumulative Gain (DCG). The DCG accumulated at a particular rank position n is defined as:

$$DCG_n = rating_1 + \sum_{i=2}^{n} \frac{rating_i}{\log_2(i)} \tag{B.1}$$

where i is the rank of the result, $rating_i$ is the graded relevance of the result at position i. The Normalized DCG is then:

$$NDCG_n = \frac{DCG_n}{DCG_{ideal_n}} \tag{B.2}$$

where DCG_{ideal_n} is the DCG value computed with the benchmark ranking at position n. We use the ranking provided by the evaluators as our benchmark ranking or gold standard.

B.2. DMOZ Classification Method

We use a popular taxonomy, such as DMOZ, as a source for applying a state of the art method that can be evaluated against the gold standard and compared to our DRR measure. Here we explain how we used the DMOZ taxonomy to infer specificity levels of entities.

The Open Directory Project[1], also called DMOZ, combines the collaborative efforts of more than 96,877 volunteers helping to categorize the Web. ODP is one of the largest and most comprehensive human-edited Web page taxonomies. It is organized as a tree-structured taxonomy with over 1,014,849 categories and more than 5.1 million sites categorized[2]. The taxonomy powers core directory services for some of the most popular portals and search engines on the Web, including AOL Search, Google, etc. and it is also used in many research projects as a large-scale and structured background knowledge. Here we use the DMOZ taxonomy to manually assess the specificity of entities by looking at their position in the DMOZ hierarchical structure. We started with the assumption that entities classified in a hierarchy in a position close to the root are less specific (broader) than entities classified in positions close to the leaves. The ODP hierarchical tree is built with one root (*Top*) connected to 16 *Top Categories* (e.g. Arts, Science, Sports, etc.) expanding then into more than 1 million categories at different depth levels. A standard state of the art approach for identifying the level of specificity of entities is to match the entities to the corresponding category in the tree structure, and then count the number of levels separating the root node and the category node (Fig. 6.3). For our experiment we tried to match the 160 random DBpedia entities of our test dataset

[1] ttp://www.dmoz.org
[2] From the ODP website, accessed January 2014

to the DMOZ taxonomy. As there is a difference between the two knowledge bases, we manually chose the closest match on DMOZ by identifying the category containing either the website clearly representing the entity, or the category with name almost identical to the entity name. Unfortunately we were able to match only 62 entities out of 160. The remaining entities had to be discarded as there was no equivalent on DMOZ or there were multiple possible categories to choose. This methodology, however, is comparable to the state of the art approaches that can automatically compute the specificity of terms using a structured and hierarchical background knowledge.

To the 62 entities, mapped on DMOZ, we gave a score starting from the level of the 16 *Top Categories*. To this was given the level number 0, all the immediate subcategories were assigned the level 1 and so on, continuing increasing 1 level for each sub-category level (as depicted in Figure 6.3). For example in the following hierarchical path: *Top* $\rightarrow$ *Science* $\rightarrow$ *Math* $\rightarrow$ *Algebra* $\rightarrow$ ..., to *Science* was assigned the value 0 and to *Algebra* the value 2. Overall, for the 62 DMOZ entities, the average value is 4.1 with maximum value 9 and minimum 0. The entities that were categorised by the five evaluators as *Specific* in our first evaluation (see previous Section) on the DMOZ hierarchy got an average value of 5.2 with standard deviation 1.5, while the *Generic* ones got an average value of 2.7 with standard deviation 1.2. According to these average values we selected our threshold for classifying the concepts as either *Generic* or *Specific*. The threshold selected is the level number 4: entities categorised with a level lower than 4 were classified as *Generic*, and with a value greater or equal to 4 were classified as *Specific*. This classification provided us 42 specific entities and 20 generic out of the total 62. This classification has been compared with the user based classification of the first evaluation and the DMOZ scores have been compared (as a specificity ranking strategy) with the user-based benchmark ranking and our Linked Data automatic rankings. More details in the following Section B.3.

B.3. Analysis of the Results

In this section we analyse the results of the two experiments conducted in order to evaluate the performance of our DRR measure compared to the gold standard. We additionally evaluate the performance of the other link-based measures (IP/OP, IP+OP, IP) and a state of the art approach using DMOZ as a background knowledge.

B.3.1. First Evaluation: Classification

On our evaluation dataset of 160 random DBpedia entities of interest we performed the classifications with the different methodologies as explained previously in this section. Afterwards, the precision of the DMOZ classification and the Linked Data measures have been computed against the manual classification performed by the 5 human evaluators. With *agreement* here we intend the number of entities classified in the same way by the two methods over the total number of entities of the dataset. In Table B.2 we show the results of this stage of evaluation.

DMOZ	DRR	IP/OP	IP+OP	IP
0.839	**0.841**	0.700	0.700	0.725

Table B.2.: First evaluation: Agreement of the different methods compared to the manual user classification.

As we can see from the results the DRR measure and the DMOZ classification have similar performance compared to the manual classification. For around 84% of the entities the two strategies classified the entities in the same way as the human evaluators. All the other LOD-based measures perform clearly worse in this classification task (around 10% worse) as they correctly match only 70/72% of the manually classified concepts. To note again that for the DMOZ method less entities are evaluated (only 62) because of the mismatch between DMOZ and DBpedia. According to these initial results our automatic measure has comparable performance with state of the art approaches such as those using a taxonomy like DMOZ as a background knowledge. The clear advantages in using Linked Data is that the background knowledge is extended on a Web scale, it is always updated with the quick evolution of the Social Web, it does not need to be pre-processed or stored and simple measures like the DRR can be computed in real-time.

B.3.2. Second Evaluation: Ranking

Despite the positive results obtained by the first evaluation, the experiment continued with a different scope: the capability of a method to rank the specificity of a set of entities. This revealed to be necessary after the feedback received by the users on the complexity of the first classification task and their need to express a more fine-grained score for the specificity. This evaluation tests the performance of our methods in ranking specificity of concepts compared to state of the art approaches and user-based rating.

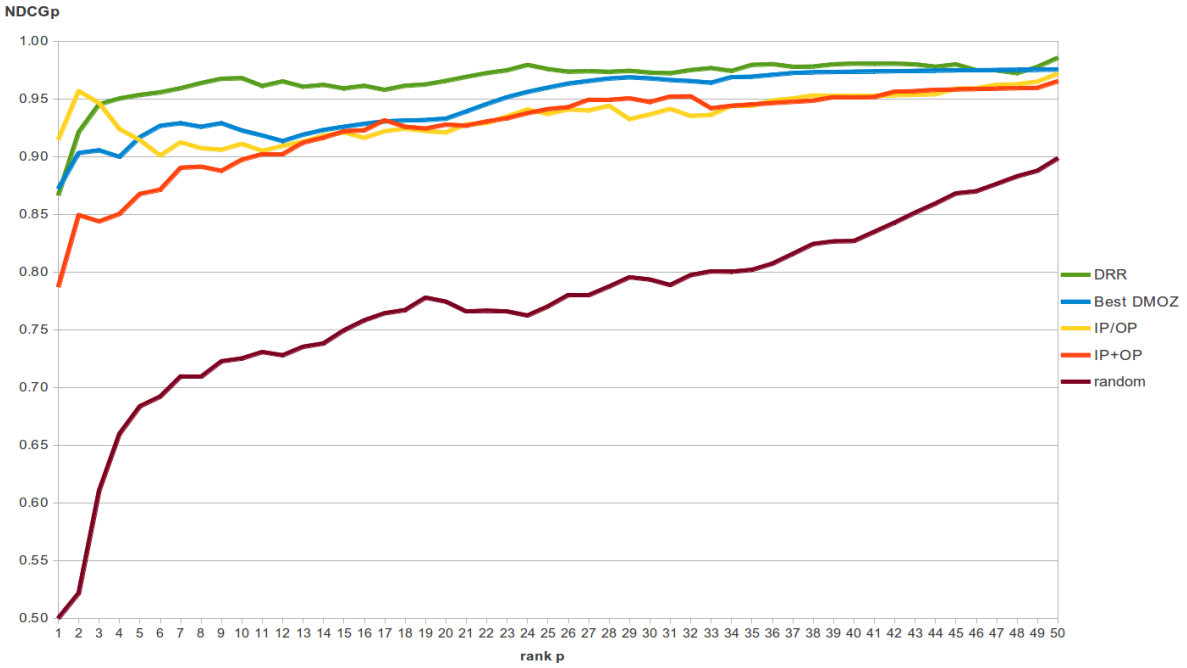

Figure B.1.: NDCG for all the different ranking positions for all the methods and 50 randomly selected entities.

For all the aforementioned methods we use the NDCG metric described in Section B.1 and we perform the ranking experiment on a subset of 50 randomly chosen entities from our complete test dataset of 160 entities. This is because of the intrinsic reduction in reliability of the NDCG measure when computed on a high rank position (effect of logarithmic reduction factor of DCG). All the NDCG measures have been computed using the human rating as gold standard (ideal ranking). In Figure B.1 we depict the NDCG (on the y-axis) computed at all the different rank positions (on the x-axis) for the 50 random entities with our different methods. In Table B.3 we summarise the NDCG values obtained for the different methods at some rank positions p.

It is clear that our DRR method that uses distinct properties is performing better than the other methods. In particular for the first 20 rank positions, where the ranking is on average almost 5% better. To note that the NDCG for the IP measure has been computed but not shown in the table as it is very close to the IP+OP one. The random method shown in the table is just a random ranking function that we evaluated as a comparison. As for the DMOZ method we had to compute the NDCG values with two different strategies. Since the DMOZ method does not provide a fine grained score to the entities but only maximum 10 possible values (unlike the other methods that are then more suitable for rankings), multiple equal scores were given to groups of entities. We

then had to rank the entities first following the DMOZ method and then, for the entities sharing the same score, rank them again according to two possible rankings given by the gold standard. Therefore, following the human ranking we were able to provide the *worst* possible DMOZ ranking (*DMOZ-*) and the *best* possible one (*DMOZ+*) for the same entities. In Figure B.1 we depict only the Best DMOZ method, but in Table B.3 we include some of the NDCG values we computed for both of them. Even in this second evaluation our proposed method for characterising the specificity of entities using the DRR measure is performing better than all the other evaluated methods.

NDCG	**DMOZ-**	**DMOZ+**	**DRR**	**IP/OP**	**IP+OP**	**random**
p=10	0.902	0.923	**0.968**	0.911	0.897	0.725
p=20	0.924	0.933	**0.966**	0.921	0.928	0.774
p=50	0.965	0.975	**0.986**	0.972	0.965	0.898

Table B.3.: Second evaluation: NDCG at different rank positions p for all methods using manual human ranking as gold standard.

Bibliography

[owl, 2004] (2004). OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004 , World Wide Web Consortium. http://www.w3.org/TR/owl-features/.

[rdf, 2004] (2004). RDF Semantics. W3C Recommendation 10 February 2004, World Wide Web Consortium. http://www.w3.org/TR/rdf-mt/.

[RDF, 2004] (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004, World Wide Web Consortium. http://www.w3.org/TR/rdf-schema/.

[Abel, 2011] Abel, F. (2011). *Contextualization, User Modeling and Personalization in the Social Web*. PhD thesis.

[Abel et al., 2011a] Abel, F., Gao, Q., Houben, G., and Tao, K. (2011a). Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In *ESWC 2011 - The Semantic Web: Research and Applications*, pages 1–15.

[Abel et al., 2011b] Abel, F., Gao, Q., Houben, G. J., and Tao, K. (2011b). Analyzing user modeling on twitter for personalized news recommendations. *UMAP*.

[Abel et al., 2011c] Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011c). Analyzing User Modeling on Twitter for Personalized News Recommendations. In *International Conference on User Modeling, Adaptation and Personalization (UMAP 2011)*, pages 1–12, Girona, Spain.

[Abel et al., 2012] Abel, F., Hauff, C., Stronkman, R., Houben, G.-J., and Tao, K. (2012). Semantics + Filtering + Search = Twitcident Exploring Information in Social Web Streams. In *Hypertext 2012*. ACM.

[Abel et al., 2010a] Abel, F., Henze, N., Herder, E., and Krause, D. (2010a). Interweav-

ing Public User Profiles on the Web. In *User Modeling, Adaptation, and Personalization*, pages 16–27. Springer.

[Abel et al., 2010b] Abel, F., Henze, N., Herder, E., and Krause, D. (2010b). Linkage, aggregation, alignment and enrichment of public user profiles with Mypes. In *Proceedings of the 6th International Conference on Semantic Systems*, pages 1–8. ACM.

[Abel et al., 2011d] Abel, F., Herder, E., Houben, G.-J., Henze, N., and Krause, D. (2011d). Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction (UMUAI), Special Issue on Personalization in Social Web Systems*, pages 1–42.

[Adler et al., 2008] Adler, B., de Alfaro, L., Pye, I., and Raman, V. (2008). Measuring author contributions to the wikipedia. In *Proceedings of WikiSym '08*. ACM.

[Aroyo and Houben, 2010] Aroyo, L. and Houben, G. (2010). User modeling and adaptive Semantic Web. *Semantic Web Journal*, 1(1):105–110.

[Artz and Gil, 2007] Artz, D. and Gil, Y. (2007). A survey of trust in computer science and the Semantic Web. *Web Semantics: Science, Services and Agents on the*, 5(2):58–71.

[Assad et al., 2007] Assad, M., Carmichael, D., Kay, J., and Kummerfeld, B. (2007). PersonisAD: Distributed, active, scrutable model framework for context-aware services. *Pervasive Computing*, pages 55–72.

[Au Yeung et al., 2008] Au Yeung, C., Liccardi, I., Lu, K., Seneviratne, O., and Berners-Lee, T. (2008). Decentralization: The Future of Online Social Networking. In *Proceedings of the W3C Workshop on the Future of Social Networking Position Papers,'08*.

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007)*, pages 715–728. Lecture Notes in Computer Science. Springer.

[Ayers and Völkel, 2008] Ayers, D. and Völkel, M. (2008). Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008, World Wide Web Consortium. http://www.w3.org/TR/cooluris/.

[Baader et al., 2010] Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F. (2010). *The Description Logic Handbook: Theory, Implemen-*

tation and Applications. Cambridge University Press.

[Barabási, 2009] Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *Science (New York, N.Y.)*, 325(5939):412–3.

[Barbier et al., 2013] Barbier, G., Feng, Z., Gundecha, P., and Liu, H. (2013). *Provenance Data in Social Media*. Morgan & Claypool.

[Barbier and Liu, 2011] Barbier, G. and Liu, H. (2011). Information provenance in social media. In *SBP11 Proceedings of the 4th international conference on social computing behavioralcultural modeling and prediction*, pages 276–283.

[Berkovsky et al., 2008] Berkovsky, S., Kuflik, T., and Ricci, F. (2008). Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3):245–286.

[Berkovsky et al., 2009] Berkovsky, S., Kuflik, T., and Ricci, F. (2009). Cross-representation mediation of user models. *User Modeling and User-Adapted Interaction*, 19(1):35–63.

[Berners-Lee, 1989] Berners-Lee, T. (1989). Information Management: A Proposal. *CERN, URL: http://www.w3.org/History/1989/proposal.html [2013-08-01]*.

[Berners-Lee, 2005] Berners-Lee, T. (2005). Tim Berners-Lee Interview at ISWC 2005.

[Berners-Lee, 2006a] Berners-Lee, T. (2006a). Linked Data. Design issues for the world wide web, World Wide Web Consortium. http://www.w3.org/DesignIssues/LinkedData.html.

[Berners-Lee, 2006b] Berners-Lee, T. (2006b). Linked Data - Design Issues.

[Berners-Lee and Cailliau, 1990] Berners-Lee, T. and Cailliau, R. (1990). WorldWideWeb: Proposal for a HyperText Project. *CERN Proposal*.

[Berners-Lee et al., 2005] Berners-Lee, T., Fielding, R., and Masinter, L. (2005). RFC 3986 - Uniform Resource Identifier (URI): Generic Syntax.

[Berners-Lee and Fischetti, 1999] Berners-Lee, T. and Fischetti, M. (1999). Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. *Harper, San Francisco*.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J. A., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

[Bernstein et al., 2010] Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S., and Chi, E. (2010). Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 303–312. ACM.

[Berrueta et al., 2007] Berrueta, D., Brickley, D., Decker, S., Fernández, S., Görn, C., Harth, A., Heath, T., Idehen, K., Kjernsmo, K., Miles, A., Passant, A., Polleres, A., Polo, L., Michael Sintek, E. U. B., and Breslin, J. G. (2007). SIOC Core Ontology Specification. W3c member submission 12 june 2007, World Wide Web Consortium.

[Berrueta et al., 2008] Berrueta, D., Fernández, S., and Shi, L. (2008). Bootstrapping the Semantic Web of Social Online Communities. In *Proceedings of Workshop on Social Web Search and Mining 7 co-located with the 17th International World Wide Web Conference*, pages 1–4.

[Beynon-Davies, 2003] Beynon-Davies, P. (2003). *Database Systems*. Palgrave Macmillan.

[Bizer and al., 2009] Bizer, C. and al., E. (2009). Linked Data – The story so far.

[Bizer et al., 2009] Bizer, C., Auer, S., Kobilarov, G., Hellmann, S., Lehmann, J., Cyganiak, R., and Becker, C. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7:154–165.

[Bizer et al., 2007] Bizer, C., Cyganiak, R., and Heath, T. (2007). How to Publish Linked Data on the Web. Technical report. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

[Bizer and Schultz, 2011] Bizer, C. and Schultz, A. (2011). Berlin SPARQL Benchmark.

[Bojars, 2009] Bojars, U. (2009). *The SIOC MEthodology for Lightweight Ontology Development*. Ph.d. thesis.

[Bojārs et al., 2008] Bojārs, U., Passant, A., Cyganiak, R., and Breslin, J. G. (2008). Weaving sioc into the web of linked data. In *Proceedings of the WWW2008 Workshop Linked Data on the Web (LDOW2008)*, volume 369 of *CEUR Workshop Proceedings*. ceur-ws.org.

[Bose and Frew, 2005] Bose, R. and Frew, J. (2005). Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys (CSUR)*, 37(1):1–28.

[Boyd and Ellison, 2008] Boyd, D. and Ellison, N. B. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

[Boyd et al., 2010] Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *43rd Hawaii International Conference on System Sciences*, HICSS '10.

[Boyd and Hargittai, 2010] Boyd, D. and Hargittai, E. (2010). Facebook privacy settings. Who cares? *First Monday*, 15(8).

[Breslin and Decker, 2007] Breslin, J. and Decker, S. (2007). The Future of Social Networks on the Internet: The Need for Semantics. *IEEE Internet Computing*, 11(6):86–90.

[Breslin and Decker, 2006] Breslin, J. G. and Decker, S. (2006). Semantic Web 2.0: Creating Social Semantic Information Spaces.

[Breslin et al., 2009] Breslin, J. G., Passant, A., and Decker, S. (2009). *The Social Semantic Web*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Brickley and Guha, 2004] Brickley, D. and Guha, R. V. (2004). RDF Vocabulary Description Language 1.0: RDF Schema.

[Brickley and Miller, 2010] Brickley, D. and Miller, L. (2010). FOAF Vocabulary Specification.

[Broekstra et al., 2002] Broekstra, J., Kampman, A., and van Harmelen, F. (2002). Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In *The Semantic Web - ISWC 2002. First International Semantic Web Conference*, volume 2342 of *Lecture Notes in Computer Science*, pages 54–68. Springer.

[Brooks et al., 2004] Brooks, C., Winter, M., Greer, J., and McCalla, G. (2004). The massive user modelling system (MUMS). In *Seventh International Conference on Intelligent Tutoring Systems*. Springer.

[Brusilovsky, 2001] Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User-Adapted Interaction (UMUAI)*, 11:87–110.

[Brusilovsky and Henze, 2007] Brusilovsky, P. and Henze, N. (2007). Open corpus adaptive educational hypermedia. In *The adaptive web: methods and strategies of web personalization*, pages 671–696. Springer, lncs edition.

[Brusilovsky et al., 2007] Brusilovsky, P., Kobsa, A., and Nejdl, W. (2007). *The adaptive web: methods and strategies of web personalization*. Springer-Verlag, lncs edition.

[Bryl et al., 2010] Bryl, V., Giuliano, C., Serafini, L., and Tymoshenko, K. (2010). Supporting natural language processing with background knowledge: coreference resolution case. In *The Semantic Web - ISWC 2010*.

[Bunge, 1977] Bunge, M. (1977). *Treatise on Basic Philosophy: Ontology I: The Furniture of the World*. Riedel, Boston.

[Burke, 2002] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and UserAdapted Interaction (UMUAI)*, 12:331–370.

[Burke, 2007] Burke, R. (2007). Hybrid Web Recommender Systems. In *Adaptive Web*, pages 377 – 408.

[Carmagnola, 2009] Carmagnola, F. (2009). Handling Semantic Heterogeneity in Interoperable Distributed User Models. *Advances in Ubiquitous User Modelling*, pages 20–36.

[Carmagnola et al., 2011] Carmagnola, F., Cena, F., and Gena, C. (2011). User model interoperability: a survey. *User Modeling and User-Adapted Interaction*, pages 1–47.

[Carmagnola and Dimitrova, 2008] Carmagnola, F. and Dimitrova, V. (2008). An Evidence-Based Approach to Handle Semantic Heterogeneity in Interoperable Distributed User Models. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 73–82. Springer.

[Carroll et al., 2005] Carroll, J., Bizer, C., Hayes, P., and Stickler, P. (2005). Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*, pages 613–622, New York, New York, USA. ACM.

[Celino et al., 2011] Celino, I., Aglio, D. D., Valle, E. D., Huang, Y., Lee, T., Park, S., and Tresp, V. (2011). Making Sense of Location-based Micro-posts Using Stream Reasoning. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts'*, pages 13–18.

[Ceolin et al., 2010] Ceolin, D., Groth, P., and van Hage, W. (2010). Calculating the Trust of Event Descriptions using Provenance. In *Proceedings Of The SWPM 2010, Workshop At The 9th International Semantic Web Conference, ISWC-2010.*, Shanghai, China.

[Champin and Passant, 2010] Champin, P. and Passant, A. (2010). SIOC in Action - Representing the Dynamics of Online Communities. In *Proceedings of the 6th International Conference on Semantic Systems (I-SEMANTICS 2010)*. ACM.

[Chen et al., 2010] Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. H. (2010). Short and tweet: experiments on recommending content from information streams. In *28th ACM Conference on Human Factors in Computing Systems (CHI 2010)*. ACM Press.

[Cosley et al., 2007] Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. (2007). SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 32–41. ACM.

[Cyganiak and Jentzsch, 2010] Cyganiak, R. and Jentzsch, A. (2010). Linking Open Data cloud diagram.

[Davidson and Freire, 2008] Davidson, S. B. and Freire, J. (2008). Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD conference*.

[De Bra et al., 2010] De Bra, P., Kobsa, A., and Chin, D. (2010). *User Modeling, Adaptation, and Personalization*. Springer.

[Dean and Schreiber, 2004] Dean, M. and Schreiber, G. (2004). OWL Web Ontology Language Reference.

[Demartini, 2007] Demartini, G. (2007). Finding experts using wikipedia. In *Proceedings of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007*, pages 33–41, Busan, South Korea. Proceedings of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007, Busan, South Korea.

[Demter et al., 2012] Demter, J., Auer, S., Martin, M., and Lehmann, J. (2012). LODStats – An Extensible Framework for High-performance Dataset Analytics. In *Proceedings of the EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer.

[Denaux et al., 2005] Denaux, R., Dimitrova, V., and Aroyo, L. (2005). Integrating open user modeling and learning content management for the semantic web. *User Modeling 2005*, pages 9–18.

[Diewald, 2012] Diewald, N. (2012). Decentralized Online Social Networks. *Handbook of Technical Communication, Handbook of Applied Linguistics 8 (HAL 8)*, 8:461–505.

[Ding and Li, 2005] Ding, Y. and Li, X. (2005). Time Weight Collaborative Filtering. In *Proceedings of the 14th ACM international conference on Information and knowledge management CIKM 05*, pages 485–492.

[DiNucci, 1999] DiNucci, D. (1999). Fragmented future. *Print*, 53(4):32.

[Dong et al., 2010] Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. (2010). Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th World Wide Web Conference WWW'10*, pages 331–340.

[Doreian and Everett,] Doreian, P. and Everett, M. *Social Networks: An International Journal of Structural Analysis*.

[Easley and Kleinberg, 2010] Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, volume 1. Cambridge University Press.

[Ehrlich and Shami, 2010] Ehrlich, K. and Shami, N. S. (2010). Microblogging Inside and Outside the Workplace. In *ICWSM '10*, pages 42–49. AAAI Press.

[Erétéo et al., 2009] Erétéo, G., Buffa, M., Gandon, F., and Corby, O. (2009). Analysis of a Real Online Social Network using Semantic Web Frameworks. In *The Semantic Web - ISWC 2009*, pages 180–195. Springer.

[Eugenio, 2000] Eugenio, B. D. (2000). On the usage of Kappa to evaluate agreement on coding tasks. *Proceedings of LREC*.

[Evermann, 2009] Evermann, J. (2009). A UML and OWL description of Bunge's upper-level ontology model. *Software and Systems Modeling*.

[Facebook, 2013] Facebook (2013). Facebook Reports Fourth Quarter and Full Year 2013 Results.

[Fallis, 2008] Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(10):1662–1674.

[Ferrucci et al., 2010] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N.,

and Welty, C. (2010). Building Watson: An Overview of the DeepQA Project.

[Fink, 2003] Fink, J. (2003). *User modeling servers: Requirements, design, and evaluation*. PhD thesis, University of Duisburg-Essen.

[Flint, 1917] Flint, L. N. (1917). *Newspaper writing in high schools, containing an outline for the use of teachers*. Pub. from the Department of journalism press in the University of Kansas.

[Gandon and Sadeh, 2004] Gandon, F. and Sadeh, N. M. (2004). Semantic web technologies to reconcile privacy and context awareness. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1:241–260.

[Gil et al., 2010] Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., and Da Silva, P. P. (2010). Provenance XG Final Report. Technical report.

[Gil et al., 2007] Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L., and Myers, J. (2007). Examining the challenges of scientific workflows. *Computer. IEEE Computer Society*, 40(12).

[Golbeck et al., 2003] Golbeck, J., Parsia, B., and Hendler, J. (2003). Trust networks on the semantic web. *Cooperative Information Agents VII*, pages 238–249.

[Governor et al., 2009] Governor, J., Hinchcliffe, D., and Nickull, D. (2009). *Web 2.0 Architectures: What entrepreneurs and information architects need to know*.

[Granovetter, 1973] Granovetter, M. S. (1973). The strenght of weak ties. *American Journal of Sociology*, 78:1360.

[Grau et al., 2008] Grau, B. C., Patel-Schneider, P., Sattler, U., Parsia, B., Motik, B., and Horrocks, I. (2008). OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6:309–322.

[Graves et al., 2007] Graves, M., Constabaris, A., and Brickley, D. (2007). FOAF: Connecting People on the Semantic Web. *Cataloging & Classification Quarterly*, 43:191–202.

[Groth, 2007] Groth, P. T. (2007). *The Origin of Data: Enabling the Determination of Provenance in Multi-institutional Scientific Systems through the Documentation of Processes*. PhD thesis.

[Gruber, 1993] Gruber, T. R. (1993). Towards Principles for the Design of Ontologies

Used for Knowledge Sharing. *Formal Ontology in Conceptual Analysis and Knowledge Representation*, 43:907–928.

[Gruber, 2007] Gruber, T. R. (2007). Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, pages 1–19.

[Guarino and Giaretta, 1995] Guarino, N. and Giaretta, P. (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. In *Towards Very Large Knowledge Bases. Knowledge Building and Knowledge Sharing*, volume 1, pages 25–32.

[Hanneman and Riddle, 2005] Hanneman, R. A. and Riddle, M. (2005). *Introduction to Social Network Methods*, volume 46.

[Hansen et al., 2006] Hansen, T., Hardie, T., and Masinter, L. (2006). Guidelines and Registration Procedures for New URI Schemes - RFC4 395.

[Harris and Seaborne, 2013] Harris, S. and Seaborne, A. (2013). SPARQL 1.1 Overview.

[Harth et al., 2007] Harth, A., Polleres, A., and Decker, S. (2007). Towards a social provenance model for the Web. In *Proceedings of the Workshop on Principles of Provenance*.

[Hartig, 2009] Hartig, O. (2009). Provenance information in the web of data. In *2nd Workshop on Linked Data on the Web (LDOW 2009) at WWW*.

[Hartig and Zhao, 2009] Hartig, O. and Zhao, J. (2009). Using web data provenance for quality assessment. In *Proc. of the Role of Semantic Web in Provenance Management at ISWC*.

[Hartig and Zhao, 2010] Hartig, O. and Zhao, J. (2010). Publishing and Consuming Provenance Metadata on the Web of Linked Data. In *Proceedings of 3rd Int. Provenance and Annotation Workshop*.

[Hausenblas et al., 2009] Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., and Ayers, D. (2009). SCOVO: Using statistics on the Web of data. In *Semantic Web in Use Track of the 6th European Semantic Web Conference (ESWC2009)*.

[Hayes, 2004] Hayes, P. (2004). RDF Semantics.

[Heckmann, 2003] Heckmann, D. (2003). Introducing situational statements as an in-

tegrating data structure for user modeling, context-awareness and resource-adaptive computing. In *ABIS2003, Karlsruhe, Germany*, pages 283–286.

[Heckmann et al., 2005a] Heckmann, D., Schwartz, T., Brandherm, B., and Kröner, A. (2005a). Decentralized user modeling with UserML and GUMO. In Dolog, P. and Vassileva, J., editors, *Decentralized, Agent Based and Social Approaches to User Modeling, Workshop DASUM-05 at 9th International Conference on User Modelling, UM2005*, pages 61–66.

[Heckmann et al., 2005b] Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., and von Wilamowitz-Moellendorff, M. (2005b). Gumo the general user model ontology. In *User Modeling 2005*, Lecture Notes on Computer Science, pages 428–432. Springer Berlin / Heidelberg.

[Heckmann et al., 2005c] Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., and von Wilamowitz-Moellendorff, M. (2005c). Gumothe general user model ontology. In *User Modeling*, pages 428–432. Springer, lncs edition.

[Heitmann, 2012] Heitmann, B. (2012). An open framework for multi-source, cross-domain personalisation with semantic interest graphs. In *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*, page 313, New York, New York, USA. ACM Press.

[Heitmann et al., 2012a] Heitmann, B., Cyganiak, R., Hayes, C., and Decker, S. (2012a). An Empirically Grounded Conceptual Architecture for Applications on the Web of Data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(1):51–60.

[Heitmann et al., 2012b] Heitmann, B., Dabrowski, M., Passant, A., Hayes, C., and Griffin, K. (2012b). Personalisation of Social Web Services in the Enterprise Using Spreading Activation for Multi-Source, Cross-Domain Recommendations. In *Proceedings of the AAAI Spring Symposium on Intelligent Web Services Meet Social Computing.*

[Heitmann et al., 2012c] Heitmann, B., Dabrowski, M., Passant, A., Hayes, C., and Griffin, K. (2012c). Personalisation of Social Web Services in the Enterprise Using Spreading Activation for Multi-Source, Cross-Domain Recommendations. In *AAAI Spring Symposium on Intelligent Web Services Meet Social Computing.*

[Hellmann et al., 2009] Hellmann, S., Stadler, C., Lehmann, J., and Auer, S. (2009).

Dbpedia Live Extraction. In *On the Move to Meaningful Internet Systems: OTM 2009*, volume Lecture No, pages 1209–1223. Springer Berlin / Heidelberg, lecture no edition.

[Herlocker et al., 2004] Herlocker, J. L., Riedl, J. T., Konstan, J. A., and Terveen, L. G. (2004). Evaluating collaborative filtering recommender systems. In *ACM Transactions on Information Systems*, volume 22, pages 5–53.

[Hoisl et al., 2007] Hoisl, B., Aigner, W., and Miksch, S. (2007). Social rewarding in wiki systems - motivating the community. In *Online Communities and Social Computing*, pages 362–371. Springer.

[Hong et al., 2011] Hong, L., Convertino, G., and Chi, E. (2011). Language Matters in Twitter: A Large Scale Study. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, volume 91, pages 518–521.

[Horrocks, 2002] Horrocks, I. (2002). DAML+OIL: a Description Logic for the Semantic Web. *IEEE Data Engineering Bulletin*, 25:4–9.

[Järvelin and Kekäläinen, 2000] Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR*.

[Java et al., 2007] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter. *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.

[Jenders et al., 2013] Jenders, M., Kasneci, G., and Naumann, F. (2013). Analyzing and predicting viral tweets. In *WWW 2013*, pages 657–664. ACM Press.

[Jentzsch et al., 2009] Jentzsch, A., Samwald, M., and Andersson, B. (2009). Linking Open Drug Data. In *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS'09)*, pages 3–6.

[Kapanipathi et al., 2011a] Kapanipathi, P., Anaya, J., Sheth, A., Slatkin, B., and Passant, A. (2011a). Privacy-Aware and Scalable Content Dissemination in Distributed Social Networks. In *ISWC 2011 - Semantic Web In Use*, volume 1380, pages 1–16.

[Kapanipathi et al., 2011b] Kapanipathi, P., Orlandi, F., Sheth, A., and Passant, A. (2011b). Personalized Filtering of the Twitter Stream. In *SPIM Workshop at ISWC 2011*, pages 6–13. CEUR-WS.

[Kaplan and Haenlein, 2010] Kaplan, A. M. and Haenlein, M. (2010). Users of the world,

unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68.

[Kärger and Siberski, 2010] Kärger, P. and Siberski, W. (2010). Guarding a Walled Garden \- Semantic Privacy Preferences for the Social Web. *The Semantic Web: Research and Applications*.

[Kietzmann et al., 2011] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251.

[Kivran-Swaine et al., 2011] Kivran-Swaine, F., Govindan, P., and Naaman, M. (2011). The Impact of Network Structure on Breaking Ties in Online Social Networks: Unfollowing on Twitter. In *CHI 2011*, pages 1–4.

[Klamma et al., 2007] Klamma, R., Cao, Y., and Spaniol, M. (2007). Watching the Blogosphere: Knowledge Sharing in the Web 2.0. In *International Conference on Weblogs and Social Media, Boulder, CO*, pages 26–28. Citeseer.

[Klyne and Carroll, 2004a] Klyne, G. and Carroll, J. J. (2004a). Resource Description Framework (RDF): Concepts and Abstract Syntax.

[Klyne and Carroll, 2004b] Klyne, G. and Carroll, J. J. (2004b). Resource Description Framework (RDF): Concepts and abstract syntax. W3C Recommendation 10 February 2004, World Wide Web Consortium. http://www.w3.org/TR/rdf-concepts/.

[Knorr-Cetina, 1997] Knorr-Cetina, K. (1997). Sociality with objects: Social relations in postsocial knowledge societies. *Theory Culture and Society*, 14(4):1–30.

[Kobilarov et al., 2009] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., and Lee, R. (2009). Media Meets Semantic Web How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web (ESWC'09): Research and Applications*, pages 723–737.

[Kobsa, 2007] Kobsa, A. (2007). Generic User Modeling Systems. *The Adaptive Web: Methods and Strategies of Web Personalization*, 4321(LNCS):136–154.

[Kobsa and Fink, 2006] Kobsa, A. and Fink, J. (2006). An LDAP-based User Modeling Server and its Evaluation. *User Modeling and User-Adapted Interaction*, 16(2):129–169.

[Kobsa, 1991] Kobsa, A. E. (1991). User Modeling and User-Adapted Interaction. The Journal of Personalization Research.

[Korfiatis et al., 2006] Korfiatis, N., Poulos, M., and Bokos, G. (2006). Evaluating authoritative sources using social networks: an insight from Wikipedia. *Online Information Review*.

[Korth and Plumbaum, 2007] Korth, A. and Plumbaum, T. (2007). A framework for ubiquitous user modeling. In *IEEE International Conference on Information Reuse and Integration, 2007. IRI 2007.*, pages 291–297. IEEE.

[Kuflik, 2008] Kuflik, T. (2008). Semantically-enhanced user models mediation: Research agenda. In *Proc. of 5th International Workshop on Ubiquitous User Modeling (UbiqUM'2008), workshop at IUI*.

[Kwak et al., 2011] Kwak, H., Chun, H., and Moon, S. (2011). Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter. In *CHI 2011*.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web - WWW '10*, page 591.

[Lebo et al., 2013] Lebo, T., Sahoo, S., and McGuinness, D. (2013). PROV-O: The PROV Ontology.

[Leuf and Cunningham, 2001] Leuf, B. and Cunningham, W. (2001). *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley Professional.

[Li et al., 2010] Li, X., Lebo, T., and McGuinness, D. L. (2010). Provenance-based Strategies to Develop Trust in Semantic Web Applications. In *The Third International Provenance and Annotation Workshop (IPAW 2010)*.

[Lih, 2004] Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. *5th International Symposium on Online Journalism*, 2004.

[Ltd., 2011] Ltd., T. S. (2011). The Changeset protocol.

[Marie et al., 2013] Marie, N., Corby, O., Gandon, F., and Ribière, M. (2013). Composite interests' exploration thanks to on-the-fly linked data spreading activation. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media - HT '13*, pages 31–40, New York, New York, USA. ACM Press.

[Marlow et al., 2006] Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, page 40. ACM.

[Mathes, 2004] Mathes, A. (2004). Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication*, pages 1–18.

[McGuinness et al., 2006] McGuinness, D. L., Zeng, H., Da Silva, P., Ding, L., Narayanan, D., and Bhaowal, M. (2006). Investigations into trust for collaborative information repositories: A Wikipedia case study. In *Models of Trust for the Web (MTW06)*. Citeseer.

[Mehta and Nejdl, 2007] Mehta, B. and Nejdl, W. (2007). Intelligent Distributed User Modelling: from Semantics to Learning. In *UbiDeUM: Proc. of the UM '07 Workshop on Ubiquitous and Decentralized User Modeling*.

[Mehta et al., 2005] Mehta, B., Niederee, C., Stewart, A., Degemmis, M., Lops, P., and Semeraro, G. (2005). Ontologically-Enriched Unified User Modeling for Cross-System Personalization. In Ardissono, Liliana and Brna, Paul and Mitrovic, A., editor, *User Modeling 2005*, Lecture Notes in Computer Science, pages 151–151. Springer Berlin / Heidelberg.

[Mendes et al., 2011] Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems (I-Semantics)*, pages 1–8.

[Michlmayr et al., 2007] Michlmayr, E., Cayzer, S., and Shabajee, P. (2007). Add-A-Tag: Learning adaptive user profiles from bookmark collections. In *1st International Conference on Weblogs and Social Media (ICWSM2007), Boulder, Colorado (USA)*.

[Mika, 2007] Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semantics*, 5(1).

[Miles and Bechhofer, 2009] Miles, A. and Bechhofer, S. (2009). SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009.

[Miller, 2011] Miller, G. (2011). Social Scientists Wade Into the Tweet Stream. *Science*, 333(6051):1814–1815.

[Missier et al., 2013] Missier, P., Belhajjame, K., and Cheney, J. (2013). The W3C PROV family of specifications for modelling provenance metadata. In *EDBT/ICDT*

'13, pages 773–776.

[Montaner et al., 2003] Montaner, M., López, B., and De La Rosa, J. (2003). A taxonomy of recommender agents on the internet. *Artificial intelligence review*, 19:285–330.

[Montgomery and Smith, 2009] Montgomery, A. L. and Smith, M. D. (2009). Prospects for Personalization on the Internet. *Journal of Interactive Marketing*, 23(2):130–137.

[Moreau, 2010] Moreau, L. (2010). The Foundations for Provenance on the Web. *Foundations and Trends in Web Science*, 2(2-3):99–241.

[Moreau et al., 2009] Moreau, L., Clifford, B., Freire, J., Gil, Y., Groth, P., Futrelle, J., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., and Others (2009). The Open Provenance ModelCore Specification (v1. 1). *Future Generation Computer Systems*.

[Moreau and Groth, 2013] Moreau, L. and Groth, P. (2013). *Provenance: An Introduction to PROV*. Morgan & Claypool.

[Moreau and Missier, 2013a] Moreau, L. and Missier, P. (2013a). PROV-DM: The PROV Data Model.

[Moreau and Missier, 2013b] Moreau, L. and Missier, P. (2013b). PROV-DM: The PROV Data Model.

[Nakatsuji and Fujiwara, 2012] Nakatsuji, M. and Fujiwara, Y. (2012). Collaborative filtering by analyzing dynamic user interests modeled by taxonomy. In *The Semantic Web ISWC 2012*, pages 1–16.

[Niederée et al., 2004] Niederée, C., Stewart, A., Mehta, B., and Hemmje, M. (2004). A Multi-Dimensional, Unified User Model for Cross-System Personalization. In *Proceedings of the AVI Workshop on Environments for Personalized Information Access*, pages 34–54.

[Noll and Meinel, 2007] Noll, M. G. and Meinel, C. (2007). Web Search Personalization via Social Bookmarking and Tagging. In *The Semantic Web*, pages 367–380.

[Nottingham and Sayre, 2005] Nottingham, M. and Sayre, R. (2005). The Atom Syndication Format.

[Nov, 2007] Nov, O. (2007). What motivates wikipedians? *Communications of the ACM*, 50(11):64.

[O'Reilly, 2005] O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business

Models for the Next Generation of Software.

[Orlandi, 2008] Orlandi, F. (2008). Using and extending the SIOC ontology for a fine-grained wiki modeling.

[Orlandi, 2012] Orlandi, F. (2012). Multi-source provenance-aware user interest profiling on the social semantic web. In *User Modeling, Adaptation, and Personalization (UMAP 2012), Doctoral Consortium*. Springer.

[Orlandi et al., 2012] Orlandi, F., Breslin, J., and Passant, A. (2012). Aggregated, interoperable and multi-domain user profiles for the social web. In *I-SEMANTICS*.

[Orlandi et al., 2010] Orlandi, F., Champin, P.-A., and Passant, A. (2010). Semantic Representation of Provenance in Wikipedia. In *Semantic Web Provenance Management workshop (SWPM2010) at ISWC2010*, volume 1380, Shanghai. CEUR-WS.

[Orlandi et al., 2013] Orlandi, F., Kapanipathi, P., Sheth, A., and Passant, A. (2013). Characterising concepts of interest leveraging Linked Data and the Social Web. In *IEEE/WIC/ACM International Conference on Web Intelligence*, number i, Atlanta, GA, USA.

[Orlandi et al., 2014] Orlandi, F., Kapanipathi, P., Sheth, A., and Passant, A. (2014). Real-time semantic personalisation of Social Web streams. In *Submitted at the ESWC 2014 conference (under review)*.

[Orlandi and Passant, 2009] Orlandi, F. and Passant, A. (2009). Enabling cross-wikis integration by extending the SIOC ontology. In *4th Semantic Wiki Workshop (SemWiki 2009)*. CEUR-WS.

[Orlandi and Passant, 2010] Orlandi, F. and Passant, A. (2010). Semantic Search on Heterogeneous Wiki Systems. In *International Symposium on Wikis (WikiSym2010)*. ACM.

[Orlandi and Passant, 2011] Orlandi, F. and Passant, A. (2011). Modelling provenance of DBpedia resources using Wikipedia contributions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):149–164.

[Passant, 2008] Passant, A. (2008). :me owl:sameAs flickr:33669349@N00 . In *LDOW 2008 workshop*, pages 0–1.

[Passant, 2010] Passant, A. (2010). Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. In *AAAI Spring Symposium: Linked Data*

Meets Artificial Intelligence, pages 93–98.

[Passant et al., 2009a] Passant, A., Kärger, P., Hausenblas, M., Olmedilla, D., Polleres, A., and Decker, S. (2009a). Enabling Trust and Privacy on the Social Web. In *W3C Workshop on the Future of Social Networking)*.

[Passant et al., 2009b] Passant, A., Laublet, P., Breslin, J. G., and Decker, S. (2009b). A URI is worth a thousand tags: from tagging to Linked Data with MOAT. *International Journal on Semantic Web and Information Systems*, 5:71–94.

[Patel-Schneider et al., 2004] Patel-Schneider, P. F., Hayes, P., and Horrocks, I. (2004). OWL Web Ontology Language Semantics and Abstract Syntax.

[Pazzani and Billsus, 2007] Pazzani, M. J. and Billsus, D. (2007). Content-Based Recommendation Systems. In *The Adaptive Web*, pages 325–341.

[Perreault, 2011] Perreault, S. (2011). RFC6351: xCard - vCard XML Representation.

[Ponzetto and Strube, 2007] Ponzetto, S. P. and Strube, M. (2007). Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212.

[Prud'hommeaux and Seaborne, 2008] Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF. *W3C Recommendation*, pages 1–106.

[Ram and Liu, 2007] Ram, S. and Liu, J. (2007). Understanding the semantics of data provenance to support active conceptual modeling. In *Active conceptual modeling of learning*, pages 17–29. Springer Berlin / Heidelberg, lncs edition.

[Ram and Liu, 2009] Ram, S. and Liu, J. (2009). A New Perspective on Semantics of Data Provenance. In *First International Workshop on the role of Semantic Web in Provenance Management (SWPM 2009)*.

[Rizzo and Troncy, 2011] Rizzo, G. and Troncy, R. (2011). NERD: Evaluating Named Entity Recognition Tools in the Web of Data. In *ISWC'11 - Workshop on Web Scale Knowledge Extraction (WEKEX'11)*, Bonn, Germany.

[Russell and Norvig, 2003] Russell, S. J. and Norvig, P. (2003). *Artificial intelligence: a modern approach*. Prentice hall Englewood Cliffs.

[Sacco and Breslin, 2012] Sacco, O. and Breslin, J. G. (2012). PPO & PPM 2.0: Extending the Privacy Preference Framework to provide finer-grained access control for

the Web of Data. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 80–87. ACM.

[Sacco et al., 2012] Sacco, O., Orlandi, F., and Passant, A. (2012). Privacy Aware and Faceted User-Profile Management Using Social Data. *Semantic Web Journal (to be published) available online*.

[Sacco et al., 2011] Sacco, O., Passant, A., and Decker, S. (2011). An Access Control Framework for the Web of Data. *2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11*, pages 456–463.

[Sarwar et al., 2001] Sarwar, B., Karypis, G., Konstan, J., and Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the tenth international conference on World Wide Web WWW 01*, 1:285–295.

[Scerri et al., 2012] Scerri, S., Cortis, K., Rivera, I., and Handschuh, S. (2012). Knowledge Discovery in distributed Social Web sharing activities. In *Making Sense of Microposts Workshop MSM2012 at WWW 2012*, pages 26–33.

[Schein et al., 2002] Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR'02*, volume 46, page 253.

[Schmidt et al., 2007] Schmidt, K., Stojanovic, L., Stojanovic, N., and Thomas, S. (2007). On enriching ajax with semantics: The web personalization use case. In *The Semantic Web: Research and Applications*, LNCS, pages 686–700. Springer.

[Schopman et al., 2010] Schopman, B., Brickly, D., Aroyo, L., Van Aart, C., Buser, V., Siebes, R., Nixon, L., Miller, L., Malaise, V., Minno, M., and Others (2010). NoTube: making the Web part of personalised TV. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, pages 1–8.

[Shakya, 2009] Shakya, A. (2009). *Creating and Sharing Structured Semantic Web Contents through the Social Web*. PhD thesis, The Graduate University for Advanced Studies (SOKENDAI) Japan.

[Shaw et al., 2009] Shaw, R., Troncy, R., and Hardman, L. (2009). Lode: Linking open descriptions of events. *ASWC2009*, 5926(Lecture Notes in Computer Science):153–167.

[Simmhan et al., 2005] Simmhan, Y., Plale, B., and Gannon, D. (2005). A survey of data provenance techniques. *Computer Science Department, Indiana University, Bloomington IN*, 47405.

[Specia and Motta, 2007] Specia, L. and Motta, E. (2007). Integrating folksonomies with the semantic web. *The Semantic Web: research and applications*.

[Stan et al., 2011] Stan, J., Maret, P., and Do, V. (2011). Semantic User Interaction Profiles for Better People Recommendation. *International Conference on Advances in Social Networks Analysis and Mining*.

[Story et al., 2009] Story, H., Harbulot, B., Jacobi, I., and Jones, M. (2009). FOAF + SSL : RESTful Authentication for the Social Web. *Semantic Web Conference*.

[Stvilia et al., 2005] Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *In Proceedings of the International Conference on Information Quality*, volume 11, pages 442—-454.

[Subramaniam et al., 2013] Subramaniam, N., Nandhakumar, J., and Baptista John, J. a. (2013). Exploring social network interactions in enterprise systems: the role of virtual co-presence. *Information Systems Journal*, 23(6):475–499.

[Szomszor et al., 2008] Szomszor, M., Alani, H., Cantador, I., OHara, K., and Shadbolt, N. (2008). Semantic modelling of user interests based on cross-folksonomy analysis. *The Semantic Web-ISWC 2008*, pages 632–648.

[Tao et al., 2011] Tao, K., Abel, F., Gao, Q., and Houben, G. (2011). TUMS: Twitter-based User Modeling Service. In *International Workshop on User Profile Data on the Social Semantic Web (UWeb), co-located with Extended Semantic Web Conference (ESWC), Heraklion, Greece*, pages 1–15.

[Tao et al., 2012] Tao, K., Abel, F., Hauff, C., and Houben, G.-J. (2012). What makes a tweet relevant for a topic? In *Making Sense of Microposts Workshop MSM2012 at WWW 2012*. CEUR-WS.

[Tapscott and Williams, 2006] Tapscott, D. and Williams, A. D. (2006). *Wikinomics: How Mass Collaboration Changes Everything*, volume 58. Portfolio.

[Theoharis et al., 2008] Theoharis, Y., Tzitzikas, Y., Kotzinos, D., and Christophides, V. (2008). On Graph Features of Semantic Web Schemas. *IEEE Transactions on Knowledge and Data Engineering*.

[Torre, 2009] Torre, I. (2009). Adaptive systems in the era of the semantic and social web, a survey. *User Modeling and User-Adapted Interaction*, 19(5):433–486.

[Van Aart et al., 2009] Van Aart, C., Aroyo, L., Raimond, Y., Brickley, D., Schreiber, G., Minno, M., Miller, L., Palmisano, D., Mostarda, M., Siebes, R., and Others (2009). The NoTube Beancounter: aggregating user data for television programme recommendation. In *Proceedings of the Linked Data on the Web Workshop (LDOW 2009), Madrid, Spain*, pages 1–12.

[Vassileva, 2001] Vassileva, J. (2001). Distributed user modelling for universal information access. *Universal access in HCI: Towards and information ...*.

[Villata et al., 2012] Villata, S., Costabello, L., Delaforge, N., and Gandon, F. (2012). A Social Semantic Web Access Control Model. *Journal on Data Semantics*, 2(1):21–36.

[Viviani et al., 2010] Viviani, M., Bennani, N., and Egyed-Zsigmond, E. (2010). A Survey on User Modeling in Multi-application Environments. In *Third International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services*, number Section II, pages 111–116. IEEE.

[Vrandecic et al., 2010] Vrandecic, D., Ratnakar, V., Krotzsch, M., and Gil, Y. (2010). Shortipedia: Aggregating and Curating Semantic Web Data. In *In Semantic Web Challenge at the International Semantic Web Conference (ISWC), Shanghai, China*, pages 1–8.

[W3C, 2012] W3C, O. W. G. (2012). OWL 2 Web Ontology Language Document Overview (Second Edition).

[Wagner, 2004] Wagner, C. (2004). Wiki: A technology for conversational knowledge management and group collaboration. *Communications of the Association for Information Systems (Volume13, 2004)*, 13(1):265–289.

[Ward and Bo, 2002] Ward, C. and Bo, L. (2002). What is wiki. http://www.wiki.org/wiki.cgi?WhatIsWiki. accessed March 2009.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, volume 8.

[Weiss, 2004] Weiss, A. (2004). Your blog? Who Gives a @*#%! *netWorker*, 8(1):40.

[Zalta, 2009] Zalta, E. N. (2009). Stanford Encyclopedia of Philosophy. *Stanford Encyclopedia of Philosophy*, 33:210–228.

[Zhou et al., 2005] Zhou, B., Hui, S., and Fong, A. (2005). Web usage mining for semantic web personalization. In *Workshop on Personalization on the Semantic Web (PerSWeb'05)*, pages 66–72, Edinburgh, Scotland.

List of Figures

List of Tables