ANALYSIS CERTIFICATE

magister by compilatio.net

Account : **Bernard Merialdo**
ID : **g3xu5sj1**
Title : **Thesis ahmad assaf-rapporteurs.pdf**
Folder : **Ahmad Assaf**
Comments : *Not available*
uploaded on the :10/07/2015 4:17 PM

Similarity document :

Similarities section 1 :

DETAILED INFORMATION

Title : Thesis Ahmad Assaf-Rapporteurs.pdf
Description : Ahmad Assaf
Analysed on : 10/07/2015 4:38 PM
Login ID : kzohx275
uploaded on the : 10/07/2015 4:17 PM
Upload type : manual submission
File name : Thesis Ahmad Assaf-Rapporteurs.pdf
File type : pdf
Word count : 14662
Character count : 76300

TOP PROBABLE SOURCES- *AMONG 5 PROBABLE SOURCES*

1. Your document: fn1679 - Thesis-Houda-Reviewers.pdf — 4%
2. Source Compilatio.net ilrz19 — 3%
3. **www.cambridgesemantics.com** — <1%

SIMILARITIES FOUND IN THIS DOCUMENT/SECTION

Matching similarities : **4 %**
Assumed similarities : **0 %**
Accidental similarities : **<1 %**

Highly probable sources - 5

Accidental sources- 2 Sources

Less probable sources - 14 Ignored sources - 11

HIGHLY PROBABLE SOURCES

5 Sources	Similarity
1. Your document: fn1679 - Thesis-Houda-Reviewers.pdf	4%
2. Source Compilatio.net ilrz19	3%
3. **www.cambridgesemantics.com**	<1%
4. **www.xml.com**/.../a/1665	<1%
5. **www.xml.com**/.../24/rdf.html	<1%

LESS PROBABLE SOURCES

14 Sources	Similarity
1. Document: afi238 - belongs to another user	<1%
2. **tomheath.com**/.../bizer-heath-berner...is-linked-data.pdf	<1%
3. Source Compilatio.net bfht89	<1%
4. Source Compilatio.net ak456	<1%
5. Source Compilatio.net aube9	<1%
6. **lod-cloud.net**/.../	<1%
7. **en.wikipedia.org**/.../wiki/Linked_Open_Data	<1%
8. **rhizomik.net**/.../html/SemanticWeb.html	<1%
9. **www.w3.org**/.../DesignIssues/LinkedData.html	<1%
10. **www.w3.org**/.../TR/ld-glossary	<1%
11. **gradschool.unc.edu**/.../guide/ordercomponents.html	<1%
12. **en.wikipedia.org**/.../wiki/Resource_Description_Framework	<1%
13. **en.wikibooks.org**/.../Semantic_Web/The_Vision	<1%
14. **www.diva-portal.org**/.../diva2:9157/FULLTEXT01.pdf	<1%

ACCIDENTAL SOURCES

2 Sources	Similarity
1. **www.cambridgesemantics.com**/.../semantic-university/rdf-101	<1%
2. Source Compilatio.net 469jt8mo	<1%

IGNORED SOURCES

11 Sources	Similarity

1. ceur-ws.org/.../Vol-1362/PROFILES2015_paper3.pdf 9%
2. ceur-ws.org/.../Vol-1362/PROFILES2015_paper1.pdf 5%
3. www.www2015.it/.../companion/p159.pdf 2%
4. 2014.eswc-conferences.org/.../files/eswc2014pd_submission_98.pdf 1%
5. github.com/.../ahmadassaf/opendata-checker 1%
6. www.eurecom.fr/.../roomba-an-extensib...ld-dataset-pro-les 1%
7. videolectures.net/.../eswc2015_assaf_roomba 1%
8. arxiv.org/.../pdf/1205.2691.pdf <1%
9. www.eurecom.fr/.../snarc-a-semantic-s...al-news-aggregator <1%
10. www.eurecom.fr/.../snarc-an-approach-...zed-social-content <1%
11. www.eurecom.fr/.../snarc-an-approach-...zed-social-content <1%

SIMILARITIES FOUND IN THIS DOCUMENT/SECTION

Matching similarities : **4 %**

Assumed similarities : **<1 %**

Accidental similarities : **<1 %**

TEXT EXTRACTED FROM THE DOCUMENT

Enabling Self-Service Data Provisioning Through Semantic Enrichment of Data

Ahmad Assaf

Main source Document: fn1679 - Thesis-Houda-Reviewers.pdf 4%
User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech

A doctoral dissertation submitted to: TELECOM ParisTech in partial fulllment of the requirements for the degree of: Doctor of Philosophy Specialty : Computer Science and Multimedia

Jury: Reviewers: ´ Prof. Philippe Cudre-Mauroux Prof. Marie Aude Aufaure Examiners: Prof. Pierre Senellart Dr. Stefan Dietze Supervisor: Dr. Rapha¨l Troncy e Dr. Aline Senart

University of Fribourg, Switzerland ´ Ecole Centrale Paris, France

-

Telecom ParisTech, France Leibniz University, Germany

- EURECOM, France - SAP, France

In the Name of God, Most Gracious, Most Merciful

Acknowledgments

Working as a PhD student in EURECOM and SAP was

a great experience that would not be achieved without the help and support of many people, who I would like

to acknowledge here. First and foremost, I would like to thank my supervisors Dr. Rapha¨l Troncy and Dr. e Aline Senart for their invaluable support and great guidance throughout my studies. I would like to express my gratitude to them for providing me with the freedom to pursue my research and the valuable feedback along the way. This work would not have been possible without their scientic knowledge, constructive advice and deep compassion.

Main source

Document: ilrz19 - Thesis Atemezing-reviewer.pdf

User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech

3%

I would like to thank my committee members, the reviewers Prof. XXX and Dr. XXXX, and furthermore the examiners Dr. XX and Dr. XXX for their precious time, shared positive insight and guidance.

I owe my deepest gratitude to my parents, Dr. Abdel Mouti Assaf and Renad Al Fahoum and to my sisters Malak, Dima and Noor for their unwavering encouragement, devotion and love and for pushing me always to be the best.

Last but not least, special thanks go to my friends

and colleagues in SAP and EURECOM for their constant friendship, moral and innite support.

Abstract

Enterprises use a wide range of heterogeneous information systems in their business activities such as Enterprise Resource Planning (ERP), Customer Relationships Management (CRM) and Supply Chain Management (SCM) systems. In addition to the large amounts of heterogeneous data produced by those systems, external data is an important resource that can be leveraged to enable taking quick and rational business decisions. Classic Business Intelligence (BI) and even the newer Agile Visualization tools focus much of their selling features on attractive and unique visualizations. Preparing data for those visualizations still remains the far most challenging task in most BI projects large and small. Self-service data provisioning aims at tackling this problem by providing intuitive datasets discovery, data acquisition and integration techniques to the end user. The goal of this thesis is to provide a framework that enables self-service data provisioning in the enterprise. This framework empowers business users to search, inspect and reuse data through semantically enriched datasets proles. Publicly available datasets contain knowledge from various domains such as encyclopedic, government, geographic, entertainment and so on. The increasing diversity of these datasets makes it dicult to annotate them with a xed number of predened tags. Moreover, manually entered tags are subjective and may not capture their essence and breadth. We propose a mechanism to automatically attach meta information to data objects by leveraging knowledge bases like DBpedia and Freebase which facilitates data search and acquisition for business users. In many knowledge bases, data entities are described with numerous properties.

Main source

Document: ilrz19 - Thesis Atemezing-reviewer.pdf

User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech

3%

However, not all properties have the same importance. Some properties are considered as keys for performing instance matching tasks while other properties are generally chosen for quickly providing a summary of the key facts attached to an entity.

Business users may want to enrich their reports with these data entities. To facilitate this, we propose a mechanism to select what properties should be used when augmenting extra columns into an existing dataset or annotating instances with semantic tags. Linked Open Data (LOD) has emerged as one of the largest collections of interlinked datasets on the web. In order to benet from this mine of data, one needs to access to descriptive information about each dataset (or metadata). This metadata enables dataset discovery, understanding, integration and maintenance. Data portals, which are datasets' access points, oer metadata represented in dierent and heterogeneous models. We rst propose a harmonized dataset model based on a systematic literature survey that enables complete metadata coverage to enable data discovery, exploration and reuse by business users. Second, rich metadata information is

Contents

vii

currently very limited to a few data portals where they are usually provided manually, thus being often incomplete and inconsistent in terms of quality. We propose a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset proles. This approach applies several techniques in order to check the validity of the metadata provided and to generate descriptive and statistical information for a particular dataset or for an entire data portal. Traditional data quality is a thoroughly researched eld with several benchmarks and frameworks to grasp its dimensions. Ensuring data quality in Linked Open Data is much more complex. It consists of structured information supported by models, ontologies and vocabularies and contains queryable endpoints and links. We propose an objective assessment framework for Linked Data quality based on quality metrics that can be automatically measured. We further present an extensible quality measurement tool implementing this framework that helps on one hand data owners to rate the quality of their datasets and get some hints on possible improvements, and on the other hand data consumers to choose their data sources from a ranked set. Finally, the Internet has created a paradigm shift in how we consume and disseminate information. Data nowadays is spread over heterogeneous silos of archived and live data. People willingly share data on social media by posting news, views, presentations, pictures and videos. We propose a service that brings relevant, live and archived information to the business user. The key advantage is an instantaneous access to complementary information without the need to search for it. Information appears when it is relevant enabling the user to focus on what is

really important.

Table of Contents

Table of Contents

Table of Contents

List of Figures

List of Tables

Listings

List of Publications

Journal

1. Ahmad Assaf , Rapha¨l Troncy and Aline Senart: Towards An Objective e Assessment Framework for Linked Data Quality. International Journal On Semantic Web and Information Systems, under review, 2015.

Conferences

1. Ahmad Assaf , Rapha¨l Troncy and Aline Senart: Automatic Validation, e Correction and Generation of Dataset Metadata - Enhancing Dataset Search and Spam Detection. In 24th International World Wide Web Conference (WWW 2015), Demo Track, May 2015, Florence, Italy. 2. Ahmad Assaf , Ghislain Atemezing, Rapha¨l Troncy and Elena Cabrio: What e are the important properties of an entity?

Comparing users and knowledge graph point of view. In 11th Extended Semantic Web Conference (ESWC 2014), Demo Track, May 2014, Heraklion, Crete.

3. Ahmad Assaf , Aline Senart and Rapha¨l Troncy: SNARC - An Approach e for Aggregating and Recommending Contextualized Social Content. In 10th Extended Semantic Web Conference (ESWC 2013), Sattelite Events, May 2013, Montpellier, France. 1st Prize Winner of the AI Mashup Challenge

Workshops

e 1. Ahmad Assaf , Rapha¨l Troncy and Aline Senart: What's up LOD Cloud - Observing The State of Linked Open Data Cloud Metadata. In 2nd Workshop on Linked Data Quality (LDQ), May 2015, Portoroz, Slovenia. 2. Ahmad Assaf , Rapha¨l Troncy and Aline Senart: HDL - Towards A Hare monized Dataset Model for Open Data Portals. In 2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data (PROFILES), May 2015, Portoroz, Slovenia. 3. Ahmad Assaf , Rapha¨l Troncy and Aline Senart: An Extensible Framee work to Validate and Build Dataset Proles. In 2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data (PROFILES), May 2015, Portoroz, Slovenia. Best paper award

Acronyms

xvii

4. Ahmad Assaf , Aline Senart and Rapha¨l Troncy: Data Quality Princie ples in the Semantic Web. In International Workshop on Data Quality Management and Semantic Technologies (DQMST), July 2012, Palermo, Italy. e 5. Ahmad Assaf , Eldad Louw, Aline Senart, Corentin Follenfant Rapha¨l Troncy and David Trastour: RUBIX: a Framework for Improving Data Integration with Linked Data. In 1st International Workshop on Open Data (WOD), June 2012, Nantes, France.

Glossary

Here are the main acronyms used in this document. The meaning of an acronym is usually indicated once, when it rst appears in the text.

AIS AMC API BI CCMS CRM CSV DI DMS DW EDW ERP ETL FOAF GA HTML HTTP IR JSON KB LD LDA LOD ML NE NER NERD NLP OBD OD OGD OLAP OLTP

Active Information Store Auto Mapping Core Application Programming Interface Business Intelligence Common Core Metadata Schema Customer Relationships Management Comma Separated Values Data Integration Data Management Systems Data Warehousing Enterprise Data Warehouse Enterprise Resource Planning Extract-Transform-Load Friend Of A Friend Genetic

Algorithm Hyper Text Markup Language Hypertext Transfer Protocol Information Retrieval

JavaScript Object Notation Knowledge Base Linked Data Latent Dirichlet Allocation Linked Open Data Machine Learning Named Entity Named Entity Recognition Named Entity Recognition and Disambiguation Natural Language Processing Open Business Data Open Data Open Government Data Online Analytical Processing Online Transaction Processing

Acronyms

xix Web Ontology Language Project Open Data Pearson Product-Moment Correlation Coecient Resource Description Framework Resource Description Framework Schema Representational State Transfer Software-as-a-Service SAP High Performance Analytic Appliance Supply Chain Management Simple Knowledge Organization System Service-Oriented Architecture Protocol and RDF Query Language Universal Resource Identier Universal Resource Locator World Wide Web Consortium Extensible Markup Language

OWL POD PPMCC RDF RDFS REST SaaS SAP HANA SCM SKOS SOA SPARQL URI URL W3C XML

Chapter 1

Introduction

“More data usually beats better algorithms” Anand Rajaraman Business Intelligence (BI) has always been about creating new insight for business by converting data into meaning that can be shared between people to drive change in the organization. One key aspect of creating meaning is to have a common shared understanding of information also known as Semantics. Classic BI and even the newer Agile Visualization tools focus much of their selling features on attractive and unique visualizations. Preparing data for those visualizations however still remains the far most challenging task in most BI projects large and small. The ultimate goal of BI is to facilitate ecient decisions while eliminating some of the IT headache. Traditionally, BI approaches have been controlled by a centralized version of truth with a wall between IT and the business. Self-service data provisioning aims at removing this wall by providing intuitive dataset discovery, acquisition and integration techniques intuitively to the end user.

1.1

Context and Motivation

Enterprises use a wide range of heterogeneous information systems in their business activities such as Enterprise Resource Planning (ERP), Customer Relationships Management (CRM) and Supply Chain Management (SCM) systems. An enterprise distributed IT landscape contains multiple systems using dierent technologies and data standards [111]. In addition to this heterogeneity, the amount of information in enterprise databases and on-line data stores expands exponentially each year. Enterprise Big Data is not big in volume only, but in the associated le formats. The information is also often stored in unstructured and unknown formats. Data integration is challenging as it requires combining data residing

at dierent sources, and providing the user with a unied

view of these data [97]. In large enterprises, it is a time and resource costly task. Various approaches have been introduced to solve this integration challenge. These approaches were primarily based on XML as the data representation syntax, Web Services to provide the data exchange protocols and Service-Oriented Architecture (SOA) as a holistic approach for distributed

2

Chapter 1. Introduction

systems architecture and communication. However, it was found that these technologies are no sucient to solve the integration problems in large enterprises [54, 55]. Recently, ontology-based data integration approaches have been suggested where ontologies are used to describe the data, queries and mappings between them [148]. A slightly dierent approach is the use of the Linked Data paradigm [21] for integrating enterprise data. Enterprises like Google and Microsoft are not only using the Linked Data integration paradigm for their information systems, but are also aiming at building enterprise knowledge bases (like the Google Knowledge Graph powered in part by Freebase1) that act as a crystallization point for their structured data. Data becomes more useful when it is open, widely available, in shareable formats and when advanced computing and analysis can yield from it. The quality and amount of structured knowledge available on the web make it now feasible for companies to mine this huge amount of public data and integrate it in their nextgeneration enterprise information management systems. An example of this external data is the Linked Open Data (LOD) cloud. From

12 datasets cataloged in 2007, it has grown today to nearly 1000 datasets containing more than 82 billion triples2 [21]. Data is being published by both the public and private sectors and covers a diverse set of domains from life sciences to media or government data. The LOD cloud is potentially a gold mine for organizations and individuals who are trying to leverage external data sources in order to produce more informed business decisions [28]. This external data can be accessed through public data portals like datahub.io and publicdata.eu or private ones like quandl.com and enigma.io. Analyzing this new type of data within the context of existing enterprise data should bring them new or more accurate business insights and allow better recognition of sales and market opportunities [94].

1.2 Use Case Scenario

To enable wide scale and ecient integration of data, there are some eorts needed from various sides. In this thesis, we tackle the issues and challenges from the point of views of two personae: • Data Analyst: A Data Analyst is an experienced professional who is able to collect and acquire data from multiple data sources, lter and clean data, interpret and analyze results and provide ongoing reports. • Data Portal Administrator: A Data Portal Administrator monitors the overall health of a portal. He oversees the creation of users, organizations and datasets. Administrators try to ensure a certain data quality level by

1 2

http://freebase.com http://datahub.io/dataset?tags=lod

continuously checking for spam and manually enhancing dataset descriptions and annotations. Throughout this thesis, we will present a use case scenario involving the two personae to illustrate the challenges and solutions that we provide. In our scenario, Dan is a Data Analyst working with the Ministry of Transport in France. His favorite tool for crunching, manipulating and visualizing data is SAP Lumira3 , a self-service data visualization tool that makes it easy to import data from multiple sources, perform visual BI analysis using intuitive dashboards, interactive maps, charts, and infographics. Dan receives a memo from his management to create a report comparing the number of car accidents that occurred in France for this year, to its counterpart in the United Kingdom (UK). In addition, he is asked to highlight accidents related to illegal consumption of alcohol in both countries. After examining the ministry's records, Dan is able to collect the data needed to create his report for the French side. Dan also issues an ocial request to the Department of Transport in UK to collect the data needed. However, Dan knows that the process takes a long time and his management needs the report within days. Dan is familiar with the Open Data movement and starts his journey searching through dierent data portals in the UK. Paul is a Data Portal Administrator for the data.gov.uk. He continuously oversees the processes of acquiring, preparing and publishing datasets. Paul always tries to ensure that the data published is of high quality and contains sucient attached metadata to easily enable search and discovery. Paul often receives complaints about inaccurate or spam datasets. He manually removes and xes errors while keeping open communication channels with the data-publishing departments.

1.3 Research Challenges

In the scenario presented above, both publishers (Data Portal Administrators) and users (Data Analysts) need pragmatic solutions that help them in their tasks. To enable that, there are some challenging research questions that have to be addressed. These challenges are organized in three main categories as the following:

1.3.1 Dataset Integration and Enrichment

• The enterprise heterogeneous data sources raise tremendous challenges. They have inherently dierent le formats, access protocols or query languages. They possess their own data model with dierent ways of representing and storing the data. Data across these sources may be noisy (e.g. duplicate or inconsistent), uncertain or semantically similar but yet dierent. Paul needs powerful

3

http://saplumira.com/

tools to map and organize the data in order to have a unied view for these heterogeneous and complex data structures. • Attaching metadata and semantic information to instances can be tricky. An entity is usually not associated with a single generic type in the knowledge base, but rather with a set of specic types which can be relevant or not given the context. Paul is challenged with nding the most relevant entity type within a given context. • Entities play a key role in knowledge bases in general and in the Web of Data in particular. Entities like those in DBpedia, are generally described with a lot of properties. However, it is dicult for Dan to assess which ones are more "important" than others for particular tasks such data augmentation and visualizing the key facts of an entity. • Social networks are not just gathering Internet users into groups of common interests, they are also helping people follow breaking news, contribute to online debates or learn from others. They are transforming Web usage in terms of users' initial entry point, search, browsing and purchasing behavior. However,

integrating information from these social networks can be tricky to Paul due to the vast amount of data available which makes hard to spot what is relevant in a timely manner.

1.3.2

Dataset Maintenance & Discovery

• Even though popular datasets like DBPedia4 and Freebase are well known and widely used, there are other hidden useful datasets not being used. Indeed these datasets may be useful for specialized domains, however without proper registry of topics, it is dicult for data analysts like Dan to nd them [91]. • The growing amount of data requires rich metadata in order to reach its full potential. This metadata enables dataset discovery, understanding, integration and maintenance. Despite the various models and vocabularies describing datasets metadata, the ability to have an overview of the dataset by inspecting its metadata can be limited. For example, Dan has diculties nding datasets with a specic geographical coverage as this information is missing from almost all of the examined datasets proles. • Users, organizations and governments are empowered to publish datasets. However, data portal administrators like Paul need to continuously and manually check portals to detect spam and maintain high quality data.

4

http://dbpedia.org

1.4. Thesis Contributions

5

1.3.3

Dataset Quality

Linked Data consists of structured information supported by models, ontologies and vocabularies and contains query endpoints and links. This makes data quality assurance a challenge. Despite the fact that Linked Open Data quality is a trending and highly demanded topic, very few eorts are currently trying to standardize, track and formalize frameworks to issue scores or certicates that will help data consumers in their integration tasks. Data portal administrators like Paul need to have an overall view of their portals quality and want to incorporate such metrics in the existing dataset proles. On the other hand, data analysts and users like Dan want to know beforehand if the dataset on hand is of a certain degree of quality to be used in their reports.

1.4

Thesis Contributions

In this thesis, we propose a framework (see Figure 1.1) to enable self-service data provisioning for internal and external data sources in the enterprise. The framework contributes to the three main challenges described above.

following tasks: • We surveyed the landscape of various models and vocabularies that describe datasets on the web. Since establishing a common vocabulary or model is the key to communication, we identied the need for an harmonized dataset metadata model containing sucient information so that consumers can easily understand and process datasets (see Section 3.1). First, we implemented a set of mappings between each properties of the surveyed models. This has lead to the design of HDL, a harmonized dataset model, that takes the best out of these models and extends them to ensure complete metadata coverage to enable data discovery, exploration and reuse (see Section 3.4). • We have analyzed the landscape of dataset proling tools and discovered various gaps (see Section 4.3). As a result, we proposed Roomba, a scalable automatic framework for extracting, validating, correcting and generating descriptive linked dataset proles (see Section 4.4). Roomba applies several techniques in order to check the validity of the metadata provided and to generate descriptive and statistical information for a particular dataset or for an entire data portal.

1.4.2

Contributions on Dataset Quality Control

Concerning our contributions on Linked Data quality assessment, we have achieved the following tasks: • We proposed a linked data quality assessment framework focusing on the data's objective metrics. We have identied a total of 64 quality indicators that were mapped when suitable to four main categories (entity, dataset, links, models) corresponding to the core Linked Data publishing principles. (see Section 5.3). • Upon surveying the landscape of data quality tools, we noticed a lack in automatic tools to check the dataset quality metrics proposed in our framework (see Section 5.4). As a result, we extended Roomba to perform a set of data quality checks on Linked datasets. Our extension covers most of the quality indicators proposed with focus on completeness, correctness, provenance and licensing (see Section 5.5).

1.4.3

Contributions on Dataset Integration and Enrichment

Regarding this aspect of our research, we have achieved the

following tasks:

1.5. Thesis Outline

7

• We created a framework called RUBIX that enables mashing-up potentially noisy enterprise data and external data. The framework leverages reference knowledge bases to annotate data with a set of semantic concepts (metadata). One of the advantages of this metadata is to enhance the matching process of heterogeneous data sources within an enterprise (see Section 7.2.2). • The metadata attached by RUBIX can be further used to enrich existing datasets. However, concepts are often represented with a large set of properties. To better recommend the top “important” properties for a concept, we reversed engineer the choices made by Google when creating knowledge graph panels and presented these choices explicitly using the Fresnel vocabulary, so that any application could read this conguration le for deciding which properties of an entity is worth to enrich (see Section 7.3). • Aggregating relevant social news is not an easy task. We provide an Application Programming Interface (API) that enables semantic social news aggregation called SNARC. We designed a sample frontend application leveraging SNARC’s capabilities to enable users to discover relevant social news instantly (see Chapter 8.

1.5

Thesis Outline

The work presented in this thesis rst describes a standard model to represent dataset proles. Then it focuses on techniques to automatically generate and validate these proles.

The rest of this manuscript is composed of two major

parts: In part I, we focus on the development of a framework that automatically validates and generates dataset proles. We highlight the extensibility of this framework and show the results of running it across various data portals. The contributions of this part have been published in [6, 8, 9, 10, 11, 12]. • Chapter 2 overviews the background of our work in data proling and quality assurance. We rst introduce the basic concepts in the Semantic Web and the important aspects related to (Linked) Open Data. Then, we describe the concepts of data proling and data quality. • Chapter 3 conducts a unique and comprehensive survey of seven metadata models: CKAN, DKAN, Public Open Data, Socrata, VoID, DCAT and Schema.org. We propose a Harmonized Dataset modeL (HDL) based on this survey. We describe use cases that show the benets of providing rich metadata to enable dataset discovery and search and spam detection.

8

Chapter 1. Introduction

• Chapter 4 emphasizes the need for tools that are able to identify various issues in this metadata and correct them automatically. We introduce Roomba, a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset proles. Afterwards, we present the results of running our framework on prominent data portals and analyze the results. We show that the overall state of Linked Data portals needs more attention. • Chapter 5 surveys the landscape of Linked Data quality tools and build upon previous eorts with focus on objective data quality measures. We further present a comprehensive objective quality framework applied to the Linked Open Data. We identify several gaps in the current tools and nd the need for a comprehensive evaluation and assessment framework for measuring quality on the dataset level. We extend Roomba to calculate 82% of the suggested datasets objective quality indicators. In part II, we focus on the challenges of external data integration in the enterprise. We focus on the development of a semantic enrichment framework and show the advantages of such enrichments in enhancing schema matching results and data enrichment.The contributions of this part have been published in [5, 7] • Chapter 6 overviews the background of our work in data integration and enrichment. We introduce the basic concepts in Business Intelligence and Data Warehousing and describe the various technologies and systems in SAP’s ecosystem. • Chapter 7 presents a framework that enables business users to semi-automatically combine potentially noisy data residing in heterogeneous silos. Semantically related data is identied and appropriate mappings are suggested to users. We also show that it is possible to reveal what are the “important” properties of entities by reverse engineering the choices made by Google when creating knowledge graph panels and by comparing users preferences obtained from a user survey. • Chapter 8 emphasizes the need for tools that are able to aggregate relevant social news to a certain context. We introduce SNARC, a semantic social news aggregation framework that leverages live rich data that social networks provide to build an interactive rich experience on the Internet.

Part I

Towards A Complete Dataset Prole

Overview of Part I

In Part I, we focus on the development of a

framework that automatically validates and generates dataset proles. We highlight the extensibility of this framework and show the results of running it against various data portals. In Chapter 2, we overview the background of our work in data proling and quality assurance. We rst introduce the basic concepts in the Semantic Web and the important aspects related

to (Linked) Open Data. Then, we describe the concepts of data proling and data quality. In Chapter 3, we conduct a unique and comprehensive survey of seven metadata models: CKAN, DKAN, Public Open Data, Socrata, VoID, DCAT and Schema.org. We propose a Harmonized Dataset modeL (HDL) based on this survey. We describe use cases that show the benets of providing rich metadata to enable dataset discovery, search and spam detection. In Chapter 4, we note the need for tools that are able to identify various issues in this metadata and correct them automatically. We introduce Roomba, a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset proles. Afterwards, we present the results of running our framework on prominent data portals and analyze the results. In Chapter 5, we survey the landscape of Linked Data quality tools and build upon previous eorts with focus on objective data quality measures. We further present a comprehensive objective quality framework applied to the Linked Open Data. We identify several gaps in the current tools and nd the need for a comprehensive evaluation and assessment framework for measuring quality on the dataset level. We extend Roomba to calculate 82% of the suggested datasets objective quality indicators.

Chapter 2

Background

2.1

Semantic Web

The web can be seen as a worldwide, distributed system of interconnected documents that humans can read, exchange and discuss. The original model behind the web can be roughly summarized as a way to publish documents represented in a standard form (e.g., HTML), containing links to other documents accessible through standard protocols (e.g., HTTP). The great advantage of the web is that it abstracts the physical storage and network layers involved in the information exchange between machines. This enables documents to appear directly connected to one another. However, in this paradigm machines are not able to achieve

Main source Document: fn1679 - Thesis-Houda-Reviewers.pdf
User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech 4%

tasks based on automated data processing such as search and query answering. To overcome this limitation, research elds such as Information Retrieval (IR), Machine Learning (ML), and Natural Language Processing (NLP) produced complex systems trying to automatically extract meaning from unstructured

data. A typical example would be search engines such as Yahoo1 and Google2 . Despite their success, there is still a semantic gap between what the machine understands and how the user perceives the data [112].

This is where Semantic Web intervenes trying to ll the

knowledge gap. In the same way that original Web abstracted away the

Main source **www.cambridgesemantics.com**/.../semantic-university/introduction-semantic-web <1%

network and physical layers, the Semantic Web abstracts away the document and application layers involved in the exchange of information. The Semantic Web connects facts, so that rather than linking

to a specic document or application, you can instead refer to a specic piece of information contained in that document or application. Berners-Lee et al. [17] provide the following denition for the Semantic Web: The

Main source Document: fn1679 - Thesis-Houda-Reviewers.pdf
User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech 4%

Semantic Web is not a separate Web but an extension of the current one, in which information is given well-dened meaning, better enabling computers and people to work in cooperation.

The word “semantic” itself implies meaning or understanding.

The fundamental dierence between Semantic Web and other data-related technologies is that the Se1 2

http://www.yahoo.com http://www.google.com

2.1. Semantic Web

13

mantic Web is concerned with the meaning and not the structure of data. This fundamental dierence engenders a completely dierent outlook on how

storing, querying, and displaying information might be approached. Some applications, such as those that refer to a large

amount of data from many dierent sources, benet enormously from this feature. What is meant by “semantic” in the Semantic Web is

not that computers are going to understand the meaning of anything, but that the logical pieces of meaning can be

mechanically manipulated by a machine to useful ends. Let us take for example, a use case where a website publishes a database about a specic product line, with descriptions and prices, while another publishes a database of product reviews. The Semantic Web standards make it easier to write an application to mesh those distributed databases together, so that a

computer could use the three data sources together to help an end-user make better purchasing decisions.

Standards facilitate building applications, especially in decentralized systems. To realize the Semantic Web vision, a series of technologies and standards have been proposed. We describe some of these standards in the following sections:

2.1.1

Resource Description Framework (RDF)

Resource Description Framework (RDF) [93] is a recommendation of the World Wide Web Consortium (W3C) that describes the Web resources. It can be seen as the data modeling language for the Semantic Web. Semantic Web resources can be anything that has an identity, they can be a person, document, image, location, etc. Each resource is assigned a Universal Resource Identier (URI) [15] which is a Unicode string to identify an abstract or physical resource. The most common type of URI is the Universal Resource Locator (URL) which is used to identify Web resources. A special case of a resource is a blank node for which no URI or literal is given. Blank nodes denote the existence of resources with specic attributes but without providing any information about their identity or reference. Resources can have atomic values named literal. They are simple strings that describe data values that do not have a separate existence. They can be plain (simple string combined with an optional language tag (e.g., “thesis”@en) or typed (string combined with a datatype URI and an optional language tag, e.g., ”0.99”^^datatypeURI). RDF reuses the XML Schema (W3C) datatypes3 which can be string, integer, oat, double or date, as dened by the XML Schema Datatype specication. RDF provides an intuitive knowledge representation using directed graphs, where the subjects and objects (resources) are the nodes and the predicates (properties) are the edges of that graph. This is referred to as an RDF Triple. Note that a property is

3

http://www.w3.org/TR/xmlschema-2

14

Chapter 2. Background

Main source Document: fn1679 - Thesis-Houda-Reviewers.pdf 4%
User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech

a specic aspect, characteristic, attribute, or relation used to describe a resource [93]. Resources can be described and linked by other set of statements forming a larger graph or a semantic network. An atomic RDF statement is a triple which is usually denoted as < s, p, o > and composed of: • Subject: the URI of a resource or a blank node which the statement refers to. • Predicate: a property of the subject and expresses the relationship between the subject and the object.

• Object: the value of the property. It can be a URI of a resource, a blank node or a literal. Figure 2.1 depicts an example of RDF graph-based representation for an address. An address is a structure that consists of dierent values such as a street, a city, a state and a zip-code.

Figure 2.1: Example of RDF representation of an address Several methods exist

Main source Document: fn1679 - Thesis-Houda-Reviewers.pdf 4%
User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech

for serializing the RDF data model. The most common format is RDF/XML. There exist other text-based formats introduced by W3C such as

Turtle4 and N-Triples5 which are easier to read than RDF/XML. RDF also contains data structures (containers and collections) that allow aggregating nodes or facts together. They are basically a syntactic sugar that will ease the process of writing code with no semantic expressiveness whatsoever.

2.1.2

RDF Schema

“It’s impossible to get everyone everywhere to agree on a

single label for every specic thing that ever was, is, or shall be”

4 5

http://www.w3.org/TeamSubmission/turtle http://www.w3.org/TR/n-triples

2.1. Semantic Web

15 Cambridge Semantics [131]

RDF is a simple and exible data model that describes resources using properties and values. Predicates in RDF are what describe and give meaning to statements. They act as a vocabulary or an ontology. An ontology is an explicit specication of a conceptualization [60]. It is a formal way to organize knowledge and terms and reect common understanding of a domain. Ontologies are typically represented as graphical relationships or networks as opposed to taxonomies which are usually presented hierarchically. Some core elements of an ontology are: • Class: denes a concept, type or collection within a specic domain. It encapsulates objects sharing some properties. For instance, in a geographical domain, the class Country is more specialized than the class Place.

• Individual: also known as instance or object and is

a member of a class. For instance, France is an instance of the class Country. • Property: is a binary relation describing how classes and individuals relate to each other. A datatype property connects instances with RDF literals while object property connects instances of two classes. For example, hasCity is an object property that can relate two instances of the class City. In order for Semantic Web applications to be able to share data, they must agree on common vocabulary. RDF doesn’t provide ways to dene those vocabularies and to

specify domain specic classes and properties.

To overcome this limitation, an extension of RDF called RDF Schema (RDFS) [29] provides a basic vocabulary to interpret RDF statements, describe taxonomies of classes and properties and dene very basic restrictions. RDFS as a modeling language allows for: 1) denition of classes and their instantiation, 2) denition of properties and restrictions and 3) denition of hierarchies for classes and properties. • Resources are instances of one or more class (rdfs:class).

Classes are organized in a hierarchy using rdfs:subClassOf property.

• Properties are assigned the class rdf:Property and are organized in a hierarchy using rdfs:subPropertyOf. • Restrictions on properties can be specied. For example, rdfs:domain to dene the class of the subject

and rdfs:range to dene the class of the object.

16

Chapter 2. Background

2.1.3

Web Ontology Language

RDFS provides basic hierarchies associated with simple restrictions. This limited expressivity triggered the need to dene an explicit formal description of concepts in complex domains. As a result, the Web Ontology Language (OWL) [59] which adds more vocabulary for describing properties and classes on top of RDF is the current markup language endorsed by W3C. It provides more relations between classes (e.g., disjointWith), logical properties (e.g., intersectionOf, sameAs) and enumerations (e.g., oneOf, allValuesFrom), among others.

2.1.4

SPARQL Query Language

Relational databases can be ecient for semantic databases. However, in practice, they are designed for a dierent type of workload. The fundamental operation of semantic databases is join, which is naturally expensive in relational databases. Given that we have our data modeled as RDF regardless of the underlying database choice, it is now possible to query and ask questions about our data in a very powerful way. Protocol and RDF Query Language (SPARQL) [126] is the standardized query language for RDF. A SPARQL query consists of a set of triples where each part (subject, predicate and/or object) can consist of variables alongside a set of conjunctions (e.g., logical “and”) or disjunctions (e.g., logical “or”). It works by matching the triples in the query with the existing RDF triples and resolving the variables.

2.1.5

Linked Data

The traditional approach of sharing data through independent silos is diminishing with the various advances in the Web. The Semantic Web envisages the availability of large amount of interlinked RDF data. Linked Data (LD) is a major milestone towards achieving this vision.

Main source Document: fn1679 - Thesis-Houda-Reviewers.pdf 4%
User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech

Formally, Linked Data has been dened as about “data published on the Web in such a way that it is machine readable, its meaning is explicitly dened, it is linked to other external datasets, and can in turn be linked

to from external datasets” [21]. Linked Data follows four main principles outlined by Tim Berners-Lee [16] to publish information on the Web, which are: 1.

Main source Document: ilrz19 - Thesis Atemezing-reviewer.pdf 3%
User: acktz1 - This document belongs to you Group: hpvx5 - Telecom ParisTech

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)2.2. Open Data17Figure 2.2: The LOD cloud as of May, 2007 4. Include links to other URIs. so that they can discover more things Linked Data is continuously evolving, started in 2007 with a dozen of datasets

(see Figure 2.2) to reach today thousands of datasets covering knowledge from various domains such as encyclopedic, government, geographic, entertainment and so on. The datasets have tripled in size from 2011 to 2014, with a signicant growth of nearly 271% [130]. The latest version published in April 2014 contains 1014 linked datasets connected by 2909 linksets (see Figure 2.5). One of the most widely used datasets is DBpedia6 . It is a structured knowledge extracted from multilingual versions of Wikipedia [23]. At the time of writing, the English version of DBpedia consists of 470 millions RDF triples that describe 4.0

million things covering a wide range of topics, and contains 45 million RDF links to several hundred external datasets.

In order to achieve the Linked Data vision, datasets should contain outbound links to other datasets. Signicant eorts try to automatically or semi-automatically generate these link to facilitate data discovery and to attach additional information.

2.2

Open Data

Open Data (OD) is the data that can be easily discovered, accessed, reused and redistributed by anyone [40]. Open data has both legal and technical dimensions. It

6

http://dbpedia.org

18

Chapter 2. Background

is placed in the public domain under liberal terms of use with minimal restrictions and is available in electronic formats that are non-proprietary and machine readable. Businesses, citizens and governments are encouraged to publish, share and reuse data. Figure 2.3 shows the Open Data ecosystem described by [62]. Each party in this ecosystem supplies dierent types of data (e.g., Open Business Data (OBD), Open Government Data (OGD)) to dierent types of stakeholders.

Figure 2.3: Open Data ecosystem Linked Open Data refers to the semantically linked, machine-readable open data. Tim Berners-Lee [16] outlined a 5 starts scheme to evaluate the availability of Linked Data as Linked Open Data: 1. Data available on the web in any format, even using PDF or image scan, but with an open license 2. Data delivered as machine-readable structured data, e.g., excel instead of image scan of a table 3. Data available in a non-proprietary format, e.g., CSV instead of Excel 4. All the above plus, data using open standards from W3C, e.g., RDF and SPARQL, to identify things and properties, so that people can point at other data 5. All the above, plus, to link data to other people’s data to provide context

2.2. Open Data

19

Open Data has major benets for citizens, businesses, societies and governments. It increases transparency and enables self-empowerment by improving the visibility of previously inaccessible information; allowing citizens to be better informed about policies, public spending and activities in the law making processes [62, 107]. Open Data is considered a gold mine for organizations which are trying to leverage external data sources in order to produce more informed business decisions [28]. Despite the legal issues surrounding Open Data licenses [78], McKinsey [107] estimates that Open Data in the health sector alone adds up over $300 billion to the economy every year. These huge benets led to a world-wide adoption of Open Data. Figure 2.4 shows the existence and support for open data initiatives, engagement with open data from outside government, legislative frameworks that support open data and the existence of training and support for data use and innovation [40]. Moreover, there are several reports and initiatives like Open Data Barometer7 , Open Data Monitor8 and

Global Open Data Index9 that aim at analyzing and monitoring the adoption of Open Data across the world. Going back to our scenario in 1.2, Open Data will help our analyst Dan in: • Having a transparent view on the data available by Ministry of Transport in France. This helps in preventing the possibility of wasting time and funds recollecting data that has been already collected by a dierent department. • Discovering complementary datasets from other sources. The benets of data transparency amplies when it is widely adopted in all other departments and agencies.The additional data enrich reports and enable better-informed, datadriven decisions. For example, by providing extra details on trac information at the time when accidents occurred, Dan could draw more accurate conclusions on the root cause of some of these accidents.

2.2.1

Open Licenses

Project Open Data 10 emphasizes the importance of datasets reusability as one of the main principles for open data. Open data should be made available under an open license. This is of high importance specially for organizations looking to integrate data for commercial use. Figure 2.5 shows the LOD cloud datasets licenses distribution. We notice that a considerable amount of datasets are still missing attached license information.

http://barometer.opendataresearch.org/ http://opendatamonitor.eu 9 http://index.okfn.org/ 10 https://project-open-data.cio.gov

8 7

Figure 2.4: Heat map of Open Government Data adoption according to the Open Data Barometer 2015 The Open Denition 11 denes a license as the legal conditions under which an item or piece of knowledge (also referred to as “work”) is made available. Domain dedications like Creative Commons Zero satisfy this denition although not technically a “license”. The Open Denition denes the following conditions for open licenses: • Allows free use of the work without any fee arrangement or compensation • Allows redistribution (on its own or as part of a collection) of the work • Allows distribution of modied work under the same license of the original • Allows any part

of the work to be freely used, distributed or modied

separately • Allows distribution of the work alongside other distinct works without placing restrictions on the additional ones • Doesn’t discriminate against any person or group • Allows use, redistribution, modication, and compilation for any purpose • Allows rights propagation to all to whom the work is distributed Despite the legal issues surrounding Linked Data licenses [78], it is still considered a gold mine for organizations who are trying to leverage external data sources in order to produce more informed business decisions [28]. In [107], the authors see the potential economic eect unfolding in education, transportation, consumer products, electricity, oil and gas, health care and consumer nance. They estimate the potential annual value enabled by Open Data in these domains to be 3 trillion US Dollars across seven domains.

11

http://opendenition.org/

Figure 2.5:

Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian

Bizer, Anja Jentzsch and Richard Cyganiak colored by licensing types. http://www.cosasbuenas.es/blog/how-o-is-lod-2015

2.3

Data Proling

The huge amount of published data makes it dicult to discover relevant datasets through traditional inspection of the raw data. Data proling is the process of creating descriptive information and collect statistics about that data. It is a cardinal activity when facing an unfamiliar dataset [100, 84]. Data proling is a vital task to monitor the quality of internal data in the enterprise. Halo BI report [74] states that nearly 40% of company’s data is found to be inaccurate. 25% of which is considered critical data. Data proles reect the importance of datasets without the need for detailed inspection of the raw data. It also helps in assessing the importance of the dataset, improving users’ ability to search and reuse part of the dataset and detecting irregularities to improve its quality. Data proling includes typically several tasks: • Metadata proling: Provides general information on the dataset (dataset description, release and latest update dates), legal information (license information, openness), practical information (access points, data dumps), etc. • Statistical proling: Provides statistical information about data types and patterns in the dataset (e.g., properties distribution, number of entities and RDF triples).

• Topical proling: Provides descriptive knowledge on the dataset content and structure. This can be in form of tags and categories used to facilitate search and reuse. • Quality proling: Discovers inconsistencies and anomalies in the data. Data

is considered of high quality if is appropriate for use and if it correctly represents the world constructs to which it refers [103]. Dataset proles are collections of data describing the internal structure of the dataset. They are presented as a set of metadata in dierent formats such as JSON, XML and RDF. The Linked Data publishing best practices [20] species that datasets should contain metadata needed to eectively understand and use them. Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource [125]. Having rich metadata helps in enabling: • Data discovery, exploration and reuse: In [147], it was found that users are facing diculties nding and reusing publicly available datasets. Metadata provides an overview of datasets making them more searchable and accessible. High quality metadata can be at times more important than the actual raw data especially when the costs of publishing and maintaining such data is high. • Organization and identication: The increasing number of datasets being published makes it hard to track, organize and present them to users eciently. Attached metadata helps in bringing similar resources together and distinguish useful links. • Archiving and preservation: There is a growing concern that digital resources will not survive in usable forms to the future [125]. Metadata can ensure resources survival and continuous accessibility by providing clear provenance information to track the lineage of digital resources and detail their physical characteristics.

2.4

Conclusion

In this chapter, we set up the grounds for the rest of this part of the thesis. We presented that basic concepts in semantic Web, open and linked data as well as data proling and its subtasks.

Chapter 3

Dataset Proles and Models

The value of Open Data is recognized when it is used. To ensure that, publishers need to enable people to nd datasets easily. Data portals are specically designed for this purpose. They make it easy for individuals and organizations to store, publish and discover datasets. Data portals (or data catalogs) are the entry points to discover published datasets. They are curated collections of datasets metadata that provide a set of complementary discovery and integration services. Data portals can be public like Datahub.io and publicdata.eu or private like quandl.com and enigma.io. Private portals harness manually curated data from various sources and expose them to users either freely or through paid plans. Similarly, in some public data portals, administrators manually review datasets information, validate, correct and attach suitable metadata information. This information is mainly in the form of predened tags such as media, geography, life sciences for organization and clustering purposes. There are several Data Management Systems (DMS) that power public data portals. CKAN1 is the world's leading open-source data portal platform powering web sites like DataHub, Europe's Public Data and the U.S Government's open data. Modeled on CKAN, DKAN2 is a standalone Drupal distribution that is used in various public data portals as well. In addition to these tradition data portals, there is a set of tools that allow exposing data directly as RESTful APIs like thedatatank.com.

3.1

Data Management Systems and Dataset Models

There are many data portals that host a large number of private and public datasets. Each portal present the data based on a model used by the underlying Data Management Software. In this section, we present the results of our landscape survey of the most common data management systems and dataset models.

1 2

http://ckan.org http://nucivic.com/dkan/

3.1.1

DCAT

The Data Catalog Vocabulary (DCAT) is a W3C recommendation that has been designed to facilitate interoperability between data catalogs published on the Web [104]. The goal behind DCAT is to increase datasets discoverability enabling applications to easily consume metadata coming from multiple sources. Moreover, the authors foresee that aggregated DCAT metadata can facilitate digital preservation and enable decentralized publishing and federated search. DCAT is an RDF vocabulary dening three main classes: dcat:Catalog, dcat:Dataset and dcat:Distribution. We are interested in both the dcat:Dataset class which is a collection of data that can be available for download in one or more formats and the dcat:Distribution class which describes the method with which one can access a dataset (e.g. an RSS feed, a REST API or a SPARQL endpoint).

3.1.2

DCAT-AP

The DCAT application prole for data portals in Europe (DCAT-AP)3 is a specialization of DCAT to describe public sector datasets in Europe. It denes a minimal set of properties that should be included in a dataset prole by specifying mandatory and optional properties. The main goal behind it is to enable cross-portal search and enhance discoverability. DCAT-AP has been promoted by the Open Data Support4 to be the standard for describing datasets and catalogs in Europe.

3.1.3

Dataset Usage Vocabulary

The Dataset Usage Vocabulary (DUV) [102] focuses on capturing the experience of using datasets. Publishers often lack feedback on how their datasets are being used and consumers lack an eective method to communicate their experiences. DUV basically aims at lling these gaps by describing consumers experiences, citations and feedback about a dataset.

3.1.4

ADMS

The Asset Description Metadata Schema (ADMS) [120] is also a prole of DCAT. It is used to semantically describe assets. An asset is broadly ded as something that can be opened and read using familiar desktop software (e.g. code lists, taxonomies, dictionaries, vocabularies) as opposed to something that needs to be processed like raw data. While DCAT is designed to facilitate interoperability between data catalogs, ADMS is focused on the assets within a catalog.

3 4

https://joinup.ec.europa.eu/asset/dcat_application_profile/description http://opendatasupport.eu

3.1. Data Management Systems and Dataset Models

25

3.1.5

VoID

VoID [25] is another RDF vocabulary designed specically to describe linked RDF datasets and to bridge the gap between data publishers and data consumers. In addition to dataset metadata, VoID describes the links between datasets. VoID dened three main classes: void:Dataset, void:Linkset and void:subset. We are specically interested in the void:Dataset concept. VoID conceptualizes a dataset with a social dimension. A VoID dataset is a collection of raw data, talking about one or more topics, originates from a certain source or process and accessible on the web.

3.1.6

CKAN