Enabling Self-Service Data Provisioning Through Semantic Enrichment of Data

Ahmad Assaf

A doctoral dissertation submitted to:

TELECOM ParisTech

in partial fulfillment of the requirements for the degree of:

Doctor of Philosophy

Specialty : Computer Science and Multimedia

Jury:

Reviewers:		
Prof. Philippe Cudré-Mauroux	-	University of Fribourg, Switzerland
Prof. Marie Aude Aufaure	-	École Centrale Paris, France
Examiners:		
Prof. Pierre Senellart	-	Telecom ParisTech, France
Dr. Stefan Dietze	-	Leibniz University, Germany
Supervisor:		
Dr. Raphaël Troncy	-	EURECOM, France
Dr. Aline Senart	-	SAP, France

Table of Contents

Chapter 1

Introduction

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 efficient 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 different technologies and data standards [38]. 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 file formats. The information is also often stored in unstructured and unknown formats.

Data integration is challenging as it requires combining data residing at different sources, and providing the user with a unified view of these data [33]. 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 systems architecture and communication. However, it was found that these technologies are no sufficient to solve the integration problems in large enterprises [18, 19]. Recently, ontology-based data integration approaches have been suggested where ontologies are used to describe the data, queries and mappings between them [50]. A slightly different approach is the use of the Linked Data paradigm [7] 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 Freebase[1]) 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

[1] http://freebase.com

available on the web make it now feasible for companies to mine this huge amount of public data and integrate it in their next-generation 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 triples[2] [7]. 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 [11]. 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 [31].

1.2 Use Case Scenario

To enable wide scale and efficient integration of data, there are some efforts 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, filter 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 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 Lumira[3], 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 official 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 different 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

[2] http://datahub.io/dataset?tags=lod
[3] http://saplumira.com/

of high quality and contains sufficient attached metadata to easily enable search and discovery. Paul often receives complaints about inaccurate or spam datasets. He manually removes and fixes 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 different file formats, access protocols or query languages. They possess their own data model with different 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 different. **Paul** needs powerful tools to map and organize the data in order to have a unified 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 specific types which can be relevant or not given the context. **Paul** is challenged with finding 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 difficult 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 DBPedia[4] 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 difficult for data analysts like **Dan** to find them [30].

[4] http://dbpedia.org

- 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 difficulties finding datasets with a specific geographical coverage as this information is missing from almost all of the examined datasets profiles.

- 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.

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 efforts are currently trying to standardize, track and formalize frameworks to issue scores or certificates 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 profiles. 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 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. In summary, the main contributions of this work are as follows:

1.4.1 Contributions on Dataset Maintenance & Discovery

Regarding this aspect of our research, we have achieved the 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 identified the need for an harmonized dataset metadata model containing sufficient information so that consumers can easily understand and process datasets. 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.

- We have analyzed the landscape of dataset profiling tools and discovered various gaps. As a result, we proposed Roomba, a scalable automatic framework for extracting, validating, correcting and generating descriptive linked dataset profiles. 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.

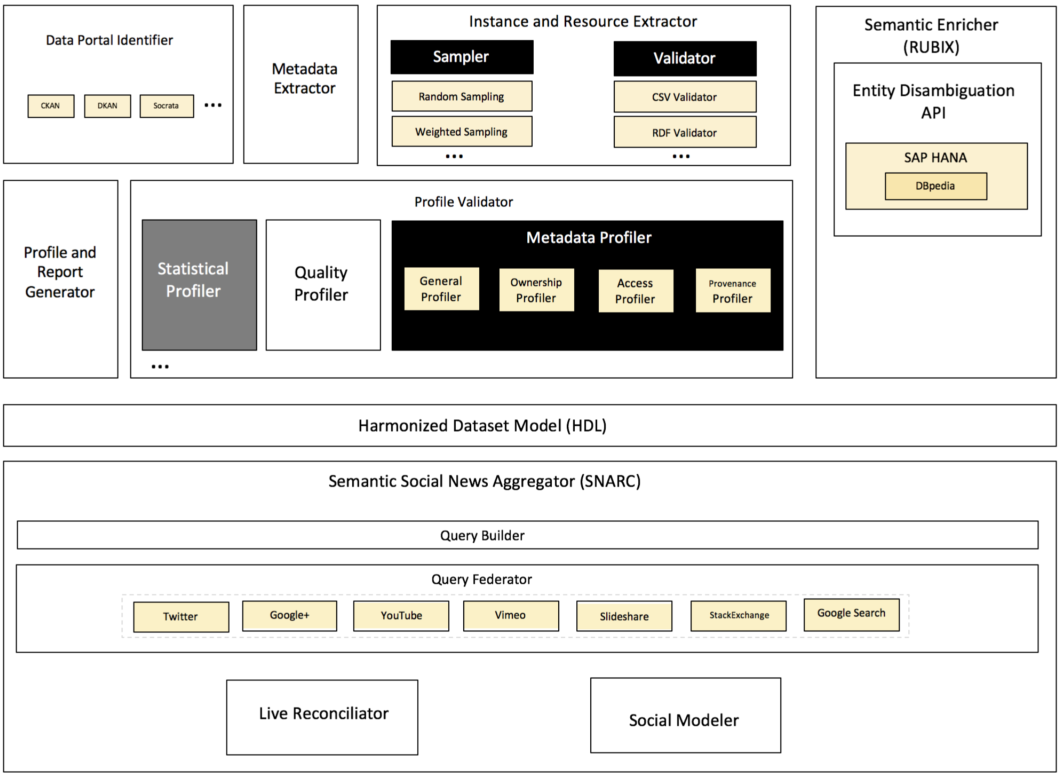

Figure 1.1: Architecture diagram for enabling self-service data provisioning

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 identified 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.

- 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. 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.

1.4.3 Contributions on Dataset Integration and Enrichment

Regarding this aspect of our research, we have achieved the following tasks:

- 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.

- 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 configuration file for deciding which properties of an entity is worth to enrich.

- 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.

Chapter 2

Towards A Complete Dataset Profile

2.1 Dataset Profiles and Models

The value of Open Data is recognized when it is used. To ensure that, publishers need to enable people to find datasets easily. Data portals are specifically 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 predefined tags such as *media, geography, life sciences* for organization and clustering purposes.

There are several Data Management Systems (DMS) that power public data portals. CKAN[1] 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, DKAN[2] 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`.

A dataset metadata model must contain sufficient information so that consumers can easily understand and process the data that is described. After analyzing the most prominent dataset models, we find out that a dataset can contain four main sections:

- **Resources**: The actual raw data that can be downloaded or accessed directly via queryable endpoints. Resources can come in various formats such as JSON, XML or RDF.
- **Tags**: Descriptive knowledge about the dataset content and structure. This can range from simple textual representation to semantically rich controlled terms. Tags are the basis for datasets search and discovery.
- **Groups**: Groups act as organizational units that share common semantics. They can be seen as a cluster or a curation of datasets based on shared categories or themes.
- **Organizations**: Organizations are another way to arrange datasets. However, they differ from groups as they are not constructed by shared semantics or properties, but solely on the dataset's association to a specific administration party.

[1] http://ckan.org

[2] http://nucivic.com/dkan/

Upon close examination of the various data models, we grouped the metadata information into eight main types. Each section discussed above should contain one or more of these types. For example, resources have general, access, ownership and provenance information while tags have general and provenance information only. The eight information types are:

- **General information**: The core information about the dataset (e.g., title, description, ID). The most common vocabulary used to describe this information is Dublin Core[3].

- **Access information**: Information about dataset access and usage (e.g., URL, license title and license URL). In addition to the properties in the models discussed above, there are several vocabularies designed specially to describe data access rights, e.g., Linked Data Rights[4], the Open Digital Rights Language (ODRL)[5].

- **Ownership information**: Authoritative information about the dataset (e.g., author, maintainer and organization). The common vocabularies used to expose ownership information are Friend-of-Friend (FOAF)[6] for people and relationships, vCard [24] for people and organizations and the Organization ontology [43] designed specifically to describe organizational structures.

- **Provenance information**: Temporal and historical information about the dataset creation and update records, in addition to versioning information (e.g., creation data, metadata update data, latest version). Provenance information coverage varies across the modeled surveyed. However, its great importance lead to the development of various special vocabularies like the Open Provenance Model[7] and PROV-O [32]. DataID [12] is an effort to provide semantically rich metadata with focus on providing detailed provenance, license and access information.

- **Geospatial information**: Information reflecting the geographical coverage of the dataset represented with coordinates or geometry polygons. There are several additional models and extensions specifically designed to express geographical information. The Infrastructure for Spatial Information in the European Community (INSPIRE) directive[8] aims at establishing an infrastructure for spatial information. Mappings have been made between DCAT-AP and the INSPIRE metadata. CKAN provides as well a spatial extension[9] to add geospatial capabilities. It allows importing geospatial metadata from other resources and supports various standards (e.g., ISO 19139) and formats (e.g., GeoJSON).

- **Temporal information**: Information reflecting the temporal coverage of the dataset (e.g., from date to date). There has been some notable work on extending CKAN to include temporal information. `govdata.de` is an Open Data portal in Germany that extends the CKAN data model to include information like `temporal_granularity`, `temporal_coverage_to` and `temporal_granularity_from`.

[3] http://dublincore.org/documents/dcmi-terms/

[4] http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/

[5] http://www.w3.org/ns/odrl/2/

[6] http://xmlns.com/foaf/spec/

[7] http://open-biomed.sourceforge.net/opmv/

[8] http://inspire.ec.europa.eu/

[9] https://github.com/ckan/ckanext-spatial

- **Statistical information**: Statistical information about the data types and patterns in datasets (e.g., properties distribution, number of entities and RDF triples). This information is particularly useful to explore a dataset as it gives detailed insights about the raw data when provided properly. VoID is the only model that provides statistical information about a dataset. VoID defines properties to express different statistical characteristics of datasets like the total number of triples, total number of entities, total number of distinct classes, etc. However, there are other vocabularies such as SCOVO [21] that can model and publish statistical data about datasets.

- **Quality information**: Information that indicates the quality of the dataset on the metadata and instance levels. In addition to that, a dataset should include an openness score that measures its alignment with the Linked Data publishing standards [5]. Quality information is only expressed in the POD metadata. However, `govdata.de` extends the CKAN model also to include a `ratings_average` field. Moreover, there are various other vocabularies like daQ [13] that can be used to express datasets quality. The RDF Review Vocabulary[10] can also be used to express reviews and ratings about the dataset or its resources.

Since establishing a common vocabulary or model is the key to communication, we identified the need for an harmonized dataset metadata model containing sufficient information so that consumers can easily understand and process datasets. To create the mappings between the different models, we performed various steps:

- Examine all the models and vocabularies specifications and documentations.

- Examine existing datasets using these models and vocabularies. Data Portals[11] provides a comprehensive list of Open Data Portals from around the world. It was our entry point to find out portals using CKAN or DKAN as their underlying DMS. We also investigated portals known to be using specific DMS. Socrata, for example, maintains a list of Open Data portals using their software on their homepage such as http://pencolorado.org and http://data.maryland.gov.

- Examine the source code of some portals. This was specifically the case for Socrata as their API returns the raw data serialized as JSON rather than the dataset's metadata. As a consequence, we had to investigate the Socrata Open Data API (SODA) source code[12] and check the different classes and interfaces.

From our survey, we found that a proper integration of Open Data into businesses requires datasets to include the following information:

- **Access information**: a dataset is useless if it does not contain accessible data dumps or query-able endpoints;

- **License information**: businesses are always concerned with the legal implications of using external content. As a result, datasets should include both machine and human readable license information that indicates permissions, copyrights and attributions;

[10] http://vocab.org/review/

[11] http://dataportals.org

[12] https://github.com/socrata/soda-java/tree/master/src/main/java/com/socrata/model

- **Provenance information**: depending on the dataset license, the data might not be legally usable if there are no information describing its authoritative and versioning information. Current models under-specify these aspects limiting the usability of many datasets.

Since establishing a common vocabulary or model is the key to communication, we identified the need for a harmonized dataset metadata model containing sufficient information so that consumers can easily understand and process datasets. We have identified four main sections that should be included in the model: resources, groups, tags and organizations. Furthermore, we have classified the information to be included into eight types. Our main contribution is a set of mappings between each properties of those models. This has lead to the design of HDL, a harmonized dataset model, that takes the best out of these models to ensure complete metadata coverage to enable data discovery, exploration and reuse.

2.2 Dataset Profiles Generation and Validation

The heterogeneous nature of data sources reflects directly on the data quality as they often contain inconsistent as well as misinterpreted and incomplete metadata information. Moreover, the significant variation in size, formats and freshness of the data, makes it more difficult to find useful datasets without prior knowledge. This can be clearly noticed in the LOD Cloud where few datasets such as DBPedia [9], Freebase [10] and YAGO [47] are favored over less popular datasets that may include domain specific knowledge more suitable for the tasks at hand. For example, for the task of building context-aware recommender systems in an academic digital library over the LOD cloud, popular datasets like the Semantic Web Dog Food[13], DBLP[14] or Yovisto[15] can be favored over lesser known but more specific datasets like VIAF[16] which links authority files of 20 national libraries, list of subject headings for public libraries in Spain[17] or the French dissertation search engine[18].

Users explore datasets in data portals relying on the metadata information attached by either the dataset owner or the data portal administrator. This information is mainly in form of predefined tags such as *media, geography, life sciences* that are used for organization and clustering purposes. However, the increasing diversity of those datasets makes it harder to classify them in a fixed number of tags that are subjectively assigned without capturing the essence and breadth of the dataset [30]. Furthermore, the increasing number of datasets available makes the manual review and curation of metadata unsustainable even when outsourced to communities.

Roomba is a tool we build to address the challenges of automatic validation and generation of descriptive datasets profiles. It is an extensible framework consisting of a processing pipeline that combines techniques for data portals identification, datasets crawling and a set of pluggable modules combining several profiling tasks. The framework validates the provided dataset metadata against an aggregated standard set of information. Metadata fields are automatically corrected when possible (e.g., adding a missing license URL reference). Moreover, a report describing all the issues that cannot be automatically fixed is created to be sent by email to the dataset's maintainer. There exist various

[13] http://datahub.io/dataset/semantic-web-dog-food
[14] http://datahub.io/dataset/dblp
[15] http://datahub.io/dataset/yovisto
[16] http://datahub.io/dataset/viaf
[17] http://datahub.io/dataset/lista-encabezamientos-materia
[18] http://datahub.io/dataset/thesesfr

statistical and topical profiling tools for both relational and Linked Data. The architecture of the framework allows to easily add them as additional profiling tasks. However, in this section, we focus on the task of dataset metadata profiling, ignoring the tasks of statistical and topical profiling. We validate our framework against a manually created set of profiles and manually check the accuracy by examining the results of running it on various CKAN-based data portals.

Roomba is built as a Command Line Interface (CLI) application using Node.js and is available on the tools Github repository[19]. Roomba allows data portal administrators like **Dan** to:

- Fetch information about the portal's data management system
- Fetch all the information about datasets from a data portal
- Fetch all the groups information from a data portal
- Crawl, fetch and cache datasets (a specific dataset, datasets in a specific group, datasets in the whole portal)
- Execute aggregation report on a specific group or on the whole data portal
- Profile a specific dataset, a whole group or the whole data portal

Figure 1.1 shows the main steps which are the following:

- **Data management system identification**: The Data Portal Identifier relies on several Web scraping techniques in the identification process which includes a combination of URL inspection, meta tags inspection and Document Object Model (DOM) inspection.
- **Metadata extraction**: After identifying the underlying portal software, The Metadata Extractor performs iterative queries to the API in order to fetch datasets metadata and persist them in a file-based cache system. Depending on the portal software, The Metadata Extractor can issue specific extraction jobs. For example, in CKAN-based portals, The Metadata Extractor is able to crawl and extract the metadata of a specific dataset, all the datasets in a specific group (e.g., LOD cloud) or all the datasets in the portal.
- **Instance and resource extraction**: From the extracted metadata, the Instance and Resource Extractor is able to identify all the resources associated with that dataset. They can have various types like a SPARQL endpoint, API, file, visualization, etc. However, before extracting the resource instance(s). Considering that certain datasets contain large amounts of resources and the limited computation power of some machines on which the framework might run on, a Sampler submodule is introduced to execute various sample-based strategies as they were found to generate accurate results even with comparably small sample size of 10% [15].
- **Profile validation**: The Profile Validator (component (iv)) identifies missing information and the ability to automatically correct them. Each set of metadata (general, access, ownership and provenance) is validated and corrected automatically when possible. Each profiler task has a set of metadata fields to check against. The validation process check if each field is defined and if the value assigned is valid.

[19] https://github.com/ahmadassaf/opendata-checker/tree/master/test

There exist many special validation steps for various fields. For example, the email addresses and URLs should be validated to ensure that the value entered is syntactically correct. In addition to that, for URLs, the Profile Validator issues an HTTP `HEAD` request in order to check if that URL is reachable. The Profile Validator also uses the information contained in a valid `content-header` response to extract, compare and correct some resources metadata values like `mimetype` and `size`.

- **Profile and report generation**: The validation process highlights the missing information and presents them in a human readable report. The report can be automatically sent to the dataset maintainer email if exists in the metadata. In addition to the generated report, the enhanced profiles are represented in JSON using the CKAN data model and are publicly available[20].

We ran our tool on two CKAN-based data portals. The first is the Datahub targeting specifically the LOD cloud group. The current state of the LOD cloud report [44] indicates that the LOD cloud contains 1014 datasets. They were harvested via an LDSpider crawler [26] seeded with 560 thousands URIs. Roomba on the other hand, fetches datasets hosted in data portals where datasets have attached relevant metadata. As a result, we relied on the information provided by the Datahub CKAN API. Examining the tags available, we found two candidate groups. The first tagged with “lodcloud” returned 259 datasets, while the second tagged with “lod” returned only 75 datasets. After manually examining the two lists, we found out the datasets grouped with the tag “lodcloud” are the correct ones as they contained more recent and accurate metadata. To qualify other CKAN-based portals for the experiments, we used `dataportals.org`, which contains a comprehensive list of Open Data portals from around the world. We chose the Amsterdam data portal [21] as it is updated frequently and highly maintained. The portal was commissioned in 2012 by the Amsterdam Economic Board Open Data Exchange (ODE), and covers a wide range of information domains (energy, economy, education, urban development, etc.) about Amsterdam metropolitan region.

In our evaluation, we focused on two aspects: i)*profiling correctness* which manually assesses the validity of the errors generated in the report, and ii)*profiling completeness* which assesses if the profilers cover all the errors in the datasets metadata.

Our evaluation showed that Roomba has complete correctness and completeness for the properties examined. As a result, we ran Roomba over the LOD cloud group hosted in the Datahub. We discovered that the general state of the examined datasets needs attention as most of them lack informative access information and their resources suffer low availability. These two metrics are of high importance for enterprises looking to integrate and use external linked data. We found out that the most erroneous information for the dataset core information are ownership related since this information is missing or undefined for 41% of the datasets. Datasets resources have the poorest metadata: 64% of the general metadata, all the access information and 80% of the provenance information contained missing or undefined values. We also showed that the automatic correction process can effectively enhance the quality of some information. We believe there is a need to have a community effort to manually correct missing important information like ownership information (maintainer, author, and maintainer and author emails).

[20] https://github.com/ahmadassaf/opendata-checker/tree/master/results

[21] http://data.amsterdamopendata.nl/

2.3 Objective Linked Data Quality Assessment

We are entering an era where open is the new default. Governments, universities, organizations and even individuals are publicly publishing huge amounts of open data. This openness should be accompanied with a certain level of trust or guarantees about the quality of data. The Linked Open Data is a gold mine for those trying to leverage external data sources in order to produce more informed business decisions [11]. However, the heterogeneous nature of sources reflects directly on the data quality as these sources often contain inconsistent as well as misinterpreted and incomplete information.

Traditional data quality is a thoroughly researched field with several benchmarks and frameworks to grasp its dimensions [27, 6, 51]. Data quality principles typically rely on many subjective indicators that are complex to measure automatically. The quality of data in indeed realized when it is used [34], thus directly relating to the ability of satisfying users' continuous needs.

Web documents that are by nature unstructured and interlinked require different quality metrics and assessment techniques than traditional datasets. For example, the importance and quality of Web documents can be subjectively calculated via algorithms like Page Rank [40]. Despite the fact that Linked Open Data quality is a trending and highly demanded topic, very few efforts are currently trying to standardize, track and formalize frameworks to issue scores or certificates that will help data consumers in their integration tasks.

Data quality assessment is the process of evaluating if a piece of data meets the consumers need in a specific use case [8]. The dimensionality of data quality makes it dependent on the task and users requirements. For example, DBpedia [9] and YAGO [47] are knowledge bases containing data extracted from structured and semi-structured sources. They are used in a variety of applications e.g., annotation systems [37], exploratory search [36] and recommendation engines [39]. However, their data is not integrated into critical systems e.g., life critical (e.g., medical applications) or safety critical (e.g., aviation applications) as its data quality is found to be insufficient.

The basic idea behind Linked Data is that its usefulness increases when it is more interlinked with other datasets. Tim Berners-Lee defined four main principles for publishing data that can ensure a certain level of uniformity reflecting directly data's usability [5]:

- **Make the data available on the Web**: assign URIs to identify things.
- **Make the data machine readable**: use HTTP URIs so that looking up these names is easy.
- **Use publishing standards**: when the lookup is done provide useful information using standards like RDF.
- **Link your data**: include links to other resources to enable users to discover more things.

Building on these principles, we group the quality attributes into four main categories:

- **Quality of the entities** : quality indicators that focus on the data at the instance level.
- **Quality of the dataset**: quality indicators at the dataset level.
- **Quality of the semantic model**: quality indicators that focus on the semantic models, vocabularies and ontologies.

- **Quality of the linking process**: quality indicators that focus on the inbound and outbound links between datasets.

In [2], the authors identified 24 different Linked Data quality attributes. These attributes are a mix of objective and subjective measures that may not be derived automatically. In this paper, we refine these attributes into a condensed framework of 10 objective measures. Since these measures are rather abstract, we should rely on quality indicators that reflect data quality [16] and use them to automate calculating datasets quality.

The quality indicators are weighted. These weights give the flexibility to define multiple degrees of importance. For example, a dataset containing people can have more than one person with the same name thus it is not always true that two entities in a dataset should not have the same preferred label. As a result, the weight for that quality indicator will be set to zero and will not affect the overall quality score for the consistency measure.

Independent indicators for entity quality are mainly subjective e.g., the degree to which all the real-world objects are represented, the scope and level of details, etc. However, since entities are governed by the underlying model, we have grouped their indicators with those of the modeling quality.

Table 2.1 lists the refined measures alongside their objective quality indicators. Those indicators have been gathered by:

- Transforming the objective quality indicators presented as a set of questions in [2] into more concrete quality indicator metrics.

- Surveying the landscape of data quality tools and frameworks.

- Examining the properties of the most prominent linked data models from the survey done in [3].

Table 2.1: Objective Linked Data quality framework

Quality Attribute	Quality Category	ID	Quality Indicator
Completeness	Dataset Level	1	Existence of supporting structured metadata [22]
		2	Supports multiple serializations [52]
		3	Has different data access points
		4	Uses datasets description vocabularies
		5	Existence of descriptions about its size
		6	Existence of descriptions about its structure (MIME Type, Format)
		7	Existence of descriptions about its organization and categorization
		8	Existence of information about the kind and number of used vocabularies [52]
	Links Level	9	Existence of dereferencable links for the dataset [22, 35, 20]
	Model Level	10	Absence of disconnected graph clusters [35]
		11	Absence of omitted top concept [22]
		12	Has complete language coverage [35]
		13	Absence of unidirectional related concepts [22]
		14	Absence of missing labels [35]
		15	Absence of missing equivalent properties [28]
		16	Absence of missing inverse relationships [28]
		17	Absence of missing domain or range values in properties [28]
Availability	Dataset Level	18	Existence of an RDF dump that can be downloaded by users [16][22]
		19	Existence of a queryable endpoint that responds to direct queries
		20	Existence of valid dereferencable URLs (respond to HTTP request)

Continued on next page

Table 2.1 Objective Linked Data quality framework

Quality Attribute	Quality Category	ID	Quality Indicator
Licensing	Dataset Level	21	Existence of human and machine readable license information [23]
		22	Existence of de-referenceable links to the full license information [23]
		23	Specifies permissions, copyrights and attributions [52]
Freshness	Dataset Level	24	Existence of timestamps that can keep track of its modifications [17]
Correctness	Dataset Level	25	Includes the correct MIME-type for the content [22]
		26	Includes the correct size for the content
		27	Absence of syntactic errors on the instance level [22]
	Links Level	28	Absence of syntactic errors [49]
		29	Use the HTTP URI scheme (avoid using URNs or DOIs) [35]
	Model Level	30	Contains marked top concepts [35]
		31	Absence of broader concepts for top concepts [35]
		32	Absence of missing or empty labels [1, 35]
		33	Absence of unprintable characters [1, 35] or extra white spaces in labels [48]
		34	Absence of incorrect data type for typed literals [22, 1]
		35	Absence of omitted or invalid languages tags [48, 35]
		36	Absence of terms without any associative or hierarchical relationships
Comprehensibility	Dataset Level	37	Existence of at least one exemplary RDF file [52]
		38	Existence of at least one exemplary SPARQL query [52]
		39	Existence of general information (title, URL, description) for the dataset
		40	Existence of a mailing list, message board or point of contact [16]
	Model Level	41	Absence of misuse of ontology annotations [35, 28]
		42	Existence of annotations for concepts [28]
		43	Existence of documentation for concepts [35, 28]
Provenance	Dataset Level	44	Existence of metadata that describes its authoritative information [17]
		45	Usage of a provenance vocabulary
		46	Usage of a versioning
Coherence	Model Level	47	Absence of misplaced or deprecated classes or properties [22]
		48	Absence of relation and mappings clashes [48]
		49	Absence of blank nodes [23]
		50	Absence of invalid inverse-functional values [22]
		51	Absence of cyclic hierarchical relations [45, 48, 35]
		52	Absence of undefined classes and properties usage [22]
		53	Absence of solely transitive related concepts [35]
		54	Absence of redefinitions of existing vocabularies [22]
		55	Absence of valueless associative relations [35]
Consistency	Model Level	56	Consistent usage of preferred labels per language tag [25, 35]
		57	Consistent usage of naming criteria for concepts [28]
		58	Absence of overlapping labels
		59	Absence of disjoint labels [35]
		60	Absence of atypical use of collections, containers and reification [22]
		61	Absence of wrong equivalent, symmetric or transitive relationships [28]
		62	Absence of membership violations for disjoint classes [22]
Security	Dataset Level	63	Uses login credentials to restrict access [52]
		64	Uses SSL or SSH to provide access to their dataset [52]

We have extended Roomba with 7 submodules that will check various dataset quality indicators. Some indicators have to be examined against a finite set. Since Roomba runs on CKAN-based data portals, we built our quality extension to calculate the scores against the CKAN standard model.

A CKAN portal contains a set of datasets $\mathbf{D} = \{D_1, ...D_n\}$. We denote the set of resources $R_i = \{r_1, ..., r_k\}$, groups $G_i = \{g_1, ..., g_k\}$ and tags $T_i = \{t_1, ..., t_k\}$ for $D_i \in \mathbf{D}(i = 1, ..., n)$ by $\mathbf{R} = \{R_1, ..., R_n\}, \mathbf{G} = \{G_1, ..., G_n\}$ and $\mathbf{T} = \{T_1, ..., t_n\}$ respectively.

Our quality framework contains a set of measures $\mathbf{M} = \{M_1, ..., M_n\}$. We denote the set of quality indicators $Q_i = \{q_1, ..., q_k\}$ for $M_i \in \mathbf{M}(i = 1, ..., n)$ by $\mathbf{Q} = \{Q_1, ..., Q_n\}$. Each quality indicator has a weight, context and a score $Q_i < weight, context, score >$. In Roomba, all the weights are equal and set to 1. However, they can be adjusted manually to rank the quality indicators. Each Q_i of M_i (for i = 1,...n) is applied to one or more of the resources, tags or groups. The indicator context is defined where $\exists Q_i \in \mathbf{R} \cup \mathbf{G} \cup \mathbf{T}$.

The quality indicator score is based on a ratio between the number of violations **V** and the total number of instances where the rule applies **T** multiplied by the specified weight for that indicator. In some cases, the quality indicator score is a boolean value (0 or 1). For example, checking if there is a valid metadata file (QI.1) or checking if the `license_url` is dereferenceable (QI.22).

$$Q \text{ weightedscore} = (V/T) * Q < weight > \tag{2.1}$$

Q weightedscore is an error ratio. A quality measure score should reflect the alignment of the dataset with respect to the quality indicators. The quality measure score **M** is calculated by dividing the weighted quality indicator scores sum by the total number of instances in its context, as the following formula shows:

$$M = 1 - ((\sum_{i=1}^{n} Q\text{i weightedscore}) / \mid Qi \text{ context} \mid) \tag{2.2}$$

Roomba covers 82% of the suggested datasets objective quality indicators. Based on our experiments running Roomba on the LOD cloud, we discovered that the general state of the datasets needs attention as most of them have low completeness, provenance, licensing and comprehensibility quality scores.

Chapter 3

Towards Enriched Enterprise Data

3.1 Data Integration in the Enterprise

Companies have traditionally performed business analysis based on transactional data stored in legacy relational databases. The enterprise data available for decision makers was typically relationship management or enterprise resource planning data. However social media feeds, weblogs, sensor data, or data published by governments or international organizations are nowadays becoming increasingly available [11].

The quality and amount of structured knowledge available make it now feasible for companies to mine this huge amount of public data and integrate it in their next-generation enterprise information management systems. 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 [31].

These new distributed sources, however, raise tremendous challenges. They have inherently different file formats, access protocols or query languages. They possess their own data model with different ways of representing and storing the data. Data across these sources may be noisy (e.g. duplicate or inconsistent), uncertain or be semantically similar yet different. Integration and provision of a unified view for these heterogeneous and complex data structures therefore require powerful tools to map and organize the data.

Establishing data knowledge bases in the enterprise can facilitate the provision of data integration services [19]. In this section, we present our work in using DBpedia as an internal knowledge base. We further present a set of services that we implemented on top of DBpedia allowing entity disambiguation and enhancing schema matching. These services enable business users to semi-automatically combine potentially noisy data residing in heterogeneous silos. Semantically related data is identified and appropriate mappings are suggested to users. On user acceptance, data is aggregated and can be visualized directly or exported to Business Intelligence reporting tools. Finally, we perform a reverse engineering of the Google Knowledge graph panel to find out what are the most relevant properties for an entity. We compare these results with a survey we conducted on 152 users and show how we can represent and explicit this knowledge using the Fresnel vocabulary.

Schema matching is typically used in business to business integration, metamodel matching, as well as ETL processes. For non-IT specialists the typical way of comparing financial data from two different years or quarters, for example, would be to copy and paste the data from one Excel spreadsheet into another one, thus creating redundancies and potentially introducing copy-and-paste errors. By using schema matching techniques it is possible to support this process semi-automatically, i.e. to determine which columns are similar and propose them to the user for integration. This integration can then be done with appropriate business intelligence tools that provide visualizations.

One of the problems in performing the integration is the quality of data. The columns may contain data that is noisy or incorrect. There may also be no column headers to provide suitable information for matching. A number of approaches exploit the similarities of headers or similarities of types of column data. We proposed a new approach that exploits semantic rich typing provided by our entity disambiguation.

3.1.1 Data Reconciliation

Reconciliation enables entity disambiguation, i.e. matching cells with corresponding typed entities in case of tabular data. Google Refine already supports reconciliation with Freebase but requires confirmation from the user. For medium to large datasets, this can be very time-consuming. To reconcile data, we therefore first identify the columns that are candidates for reconciliation by skipping the columns containing numerical values or dates. We then use the disambiguation API to query for each cell of the source and target columns the list of typed entities candidates. Results are cached in order to be retrieved by our similarity algorithms.

3.1.2 Matching Unnamed and Untyped Columns

The AMC has the ability to combine the results of different matching algorithms. Its default built-in matching algorithms work on column headers and produce an overall similarity score between the compared schema elements. It has been proven that combining different algorithms greatly increases the quality of matching results [41][46]. However, when headers are missing or ambiguous, the AMC can only exploit domain intersection and inclusion algorithms based on column data. We have therefore implemented three new similarity algorithms that leverage the rich types retrieved from Linked Data in order to enhance the matching results of unnamed or untyped columns. They are presented below.

- **Cosine Similarity**: We compare the result vector of candidate types from the source column with the result vector of candidate types from the target column. The similarity s between the columns pair can be calculated using the absolute value of the cosine similarity function.

- **Pearson Product-Moment Correlation Coefficient (PPMCC)**: The second algorithm that we implemented is PPMCC, a statistical measure of the linear independence between two variables (x, y) [29]. The input for PPMC consists of two arrays that represent the values from the source and target columns, where the source column is the column with the largest set of rich types found.

- **Spearman's Rank Correlation Coefficient**: The last algorithm that we implemented to match unnamed and untyped columns is Spearman's rank correlation coefficient. It applies a rank transformation on the input data and computes PPMCC afterwards on the ranked data. In our experiments we used Natural Ranking with default strategies for handling ties and NaN values. The ranking algorithm is however configurable and can be enhanced by using more sophisticated measures.

3.1.3 Column Labeling

We showed in the previous section how to match unnamed and untyped columns. Column labeling is however beneficial as the results of our previous algorithms can be combined with traditional header matching techniques to improve the quality of matching.

Rich types retrieved from Freebase are independent from each other. We need to find a method that will determine normalized score for each type in the set by balancing the proportion of high scores with the lower ones using Wilson score interval for a Bernoulli parameter.

3.1.4 Handling Non-String Values

So far, we have covered several methods to identify the similarity between "String" values, but how about other numeral values such as dates, money, distance, etc. For this purpose, we have implemented some basic type identifier that can recognize dates, money, numerical values, numerals used as identifiers. This will help us in better match corresponding entries. Adjusting AMC's combination algorithms can be of great importance at this stage. For example, assigning weights to different matchers and tweaking the configuration can yield more accurate results.

3.1.5 Important Properties for Entities

Entities are generally described with a lot of properties. 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. In contrast to entities, it is difficult to assess which properties are more "important".

Web scraping is a technique for extracting data from Web pages. We aim at capturing the properties depicted in the Google Knowledge Panel (GKP) that are injected in search result pages [4]. We have developed a Node.js application that queries all DBpedia concepts that have at least one instance which is `owl:sameAs` with a Freebase resource (since Freebase is the knowledge base behind the graph panel) in order to increase the probability that the search engine result page (SERP) for this resource will contain a GKP. We assume in our experiments that the properties displayed for an entity are type and context dependent (country, time, query) which can affect the results. Moreover, we filter out generic concepts by excluding those who are direct subclasses of `owl:Thing` since they will trigger ambiguous queries. We obtained a list of 352 concepts[1].

Fresnel[2] is a presentation vocabulary for displaying RDF data. It specifies *what* information contained in an RDF graph should be presented with the core concept `fresnel:Lens` [42].PROV-O[3] is a vocabulary to describe semantically rich metadata with focus on providing detailed provenance, license and access information. We use those two vocabularies to explicitly represent what properties should be depicted when displaying an entity[4]. This dataset can now be re-used as a configuration for any consuming application for a snippet of the generated Fresnel file).

3.2 Semantic Social News Aggregation

With the rapid advances of the Internet, social media become more and more intertwined with our daily lives. The ubiquitous nature of Web-enabled devices, especially mobile phones, enables users to participate and interact in many different forms like photo and video sharing platforms, forums,

[1] https://github.com/ahmadassaf/KBE/blob/master/results/dbpediaConcepts.json

[2] http://www.w3.org/2005/04/fresnel-info/

[3] http://www.w3.org/TR/prov-o/

[4] https://github.com/ahmadassaf/KBE/blob/master/results/results.n3

Algorithm 1 Google Knowledge Panel reverse engineering algorithm

```
INITIALIZE $equivalentClasses(DBpedia, Freebase)$ AS $vectorClasses$
Upload $vectorClasses$ for querying processing
Set $n$ AS number-of-instances-to-query
for each $conceptType \in vectorClasses$ do
    SELECT $n$ instances
    $listInstances \leftarrow$ SELECT-SPARQL($conceptType$, $n$)
    for each $instance \in listInstances$ do
        CALL http://www.google.com/search?q=$instance$
        if $knowledgePanel$ exists then
            SCRAP GOOGLE KNOWLEDGE PANEL
        else
            CALL http://www.google.com/search?q=$instance + conceptType$
            SCRAP GOOGLE KNOWLEDGE PANEL
        end if
        $gkpProperties \leftarrow$ GetData(DOM, EXIST(GKP))
    end for
    COMPUTE occurrences for each $prop \in gkpProperties$
end for
$gkpProperties$
```

newsgroups, blogs, micro-blogs, bookmarking services, and location-based services. 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 [14].

A common scenario that often happens while reading an interesting article, coming across a nice video or participating in a discussion in a forum is the growing interest to check related material around the information read. To do so, users might go to Twitter, Google+ or YouTube. They can try several times with several keywords to obtain the desired results. In the end, they might end up with several browser tabs opened and get distracted by the information overload from all these resources. The same happens in companies when business users are interested in information provided by corporate web applications like enterprise communities. In this section, we present SNARC, a semantic social news aggregator that leverages live rich data that social networks provide to build an interactive rich experience on both the Internet and Intranets. The service retrieves news related to the current page from popular platforms like Twitter, Google+, YouTube, Vimeo, Slideshare, StackExchange and the Web. As a possible front-end implementation, we have created a Google Chrome extension which enriches the user experience by augmenting related contextual information to entities on the page itself, as well as displaying related social news on a floating sidebar.

The back-end of SNARC consists of three major components: a document handler that creates a "Semantic Model" representing any web resource, a query layer that is responsible for disseminating queries to the supported social services and a data parser which processes the search results, wraps them in a common social model and generates the desired output.

Chapter 4

Achievements

This thesis thoroughly describes the different steps aiming at realizing the vision of enabling self service data provisioning in the enterprise. The work presented is beneficial to both our personae introduced. The contributions made are:

4.1 Contributions for Data Portals Administrators

Our data portal administrator **Paul** is always looking to expand his portals in terms of the number of datasets hosted, without compromising in their portal's data quality. In Section 2.1, we surveyed the landscape of various models and vocabularies that described datasets on the web. We found a shortcoming when it comes to having a complete descriptive dataset model taking into account access, license and provenance information. As a result, we proposed a Harmonized Dataset Model (HDL) that **Paul** will use as a basis to extend and present the datasets he controls. **Paul** now also knows what are the major dataset models out there, and what kind of metadata data owners need to fully represent their dataset. The mappings proposed will allow him to easily integrate data from various data management systems into his own.

In Section 2.2, we proposed Roomba, an automatic dataset profiles generation and validation tool that can be easily extended to perform various profiling tasks. Out of the box, **Paul** can use Roomba to automatically fix datasets metadata issues, and notify the datasets owners of the other issues to be manually fixed.

In Section 2.3, we proposed a comprehensive objective quality framework applied to the Linked Open Data. Moreover, after surveying the landscape of existing data quality tools, we identified several gaps and the need for a comprehensive evaluation and assessment framework and specifically for measuring quality on the dataset level. As a result, we presented an extension of Roomba that covers 82% of the suggested datasets objective quality indicators. **Paul** will be able now to identify spam and low quality datasets. In addition to that, data available in his portal will now have rich semantic information attached to it. For example, temporal and spatial information extracted will be assigned into the corresponding fields in HDL. As an exemplary result, various datasets will be easily identifiable to cover various parts of the UK.

4.2 Contributions for Data Analysts

Our data analyst **Dan** believes that "more data beats better algorithms" and is always hunting for high quality data to produce accurate reports to the management team. By examining the rich datasets metadata presented in HDL he will be able to make fast decisions whether the dataset examined is suitable or not. He will also have vital information about the licensing and limitations for using this data internally. He will also have assurances on the dataset quality, which will help choose the best candidates out of ranked list.

Dan will be able to have direct access to rich and high-quality dataset descriptions generated by Roomba. Moreover, the topical profilers in Roomba will be able to identify occurrences of alcohol related terms like “wine” in various datasets. Query expansion methods can be used to relate alcohol to wine allowing him to find the datasets he wants.

In Section 3.1, we presented an entity disambiguation API built on top of SAP HANA. This API is used in RUBIX, a framework we proposed to enable mashup of potentially noisy enterprise and external data. **Dan** now has access to various datasets that he found matching his query to the portal administered by **Paul**. He will be also able to use the schema matching services to find and merge those datasets in his reports.

Having imported those dataset into Lumira, he will be also able to use the internal knowledge base to apply various semantic enrichments on this data.

In Section 3.2, we proposed SNARC, a semantic social news aggregation service that allows the user to explore relevant news from internal or external sources. **Dan** is also a modern person, who is always trying to fresh information and believes in the wisdom of the crowd. Having SNARC services integrated with Lumira, he is also able to see a feed of relevant social media items that can be of interest to him. He actually follows a link in some tweet that he saw and was able to find relevant pieces of pointers that he would like to investigate further.

In summary, the contributions above pave the way to build a set of smart services to enable analysts easily find relevant pieces of information and administrators fight spam and be able to maintain high quality data portals. The work presented in this thesis goes beyond the fact that attaching metadata to datasets is vital, but propose a set of services that can automatically achieve that in seamless manner.

4.3 Perspectives

This thesis could be extended in the following directions:

4.3.1 Data Profile Representation

The proposed Harmonized Dataset Model (HDL) is currently available as a hierarchical JSON file. An enhancement would be to refine HDL and present it as a fully fledged OWL ontology. In addition, HDL can be extended to propose also a set of enumerations as values to ensure a unified fine-grained representation of a dataset. Moreover, while we presented the mappings between various models in a table structure, presenting those mappings in a machine readable format will allow various tools like Roomba to use it.

4.3.2 Automatic Dataset Profiling

It has been noticed that the issues surrounding metadata quality affect directly dataset search as data portals rely on such information to power their search index. There are various extensions to our tool Roomba that can help in automatically building and enhancing dataset profiles. An example of these extension would be the integration of statistical and topical profilers allowing the generation of full comprehensive profiles. We would also like to extend Roomba to be able to run over other data portal types like DKAN or Socrata. This extension can be done by leveraging the data models mappings we proposed. In addition to all that, a possible enhancement will be ability to correct the rest of the metadata either automatically or through intuitive manually-driven interfaces.

4.3.3 Objective Linked Data Quality

Ensuring data quality in Linked Open Data is a complex process as it consists of structured information supported by models, ontologies and vocabularies and contains queryable endpoints and links. In this thesis, we managed to narrow down the set of quality issues surrounding Linked Data to those who can be objectively measured and assessed by automatic tools. Our proposed tool covers 85% of the quality indicators proposed. A possible extension would be to integrate tools assessing models quality in addition to syntactic checkers with Roomba. This will provide a complete coverage of the proposed quality indicators. Moreover, there are currently no weights assigned to the quality indicators. A valid contribution would be to suggest weights to those indicators which will result in a more objective quality calculation process.

4.3.4 Enterprise Data Integration

A vital component to Data Integration in the enterprise is the existence of enterprise knowledge bases. Integrating additional linked open data sources of semantic types such as YAGO and evaluate our matching results against instance-based ontology alignment benchmarks such as OAEI[1] or ISLab[2] are possible future directions. Moreover, our work can be generalized to data classification. The same way the AMC helps identifying the best matches for two datasets, we plan to use it for identifying the best statistical classifiers for a sole dataset, based on normalized scores.

[1] http://oaei.ontologymatching.org/2011/instance/index.html

[2] http://islab.dico.unimi.it/iimb/

Bibliography

[1] Maribel Acosta, Amrapali Zaveri, Elena Simperl, and Dimitris Kontokostas. Crowdsourcing Linked Data quality assessment. In *12th International Semantic Web Conference (ISWC)*, 2013.

[2] Ahmad Assaf and Aline Senart. Data Quality Principles in the Semantic Web. In *6th International Conference on Semantic Computing ICSC '12*, 2012.

[3] Ahmad Assaf, Raphaël Troncy, and Aline Senart. HDL-Towards a Harmonized Dataset Model for Open Data Portals. In *2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data*, Portoroz, Slovenia, 2015.

[4] Mike Bergman. Deconstructing the Google Knowledge Graph. http://www.mkbergman.com/1009/deconstructing-the-google-knowledge-graph.

[5] Tim Berners-Lee. Linked Data - Design Issues. W3C Personal Notes, 2006. http://www.w3.org/DesignIssues/LinkedData.

[6] Gasser Les Stvilia Besiki, , Michael B. Twidale, and Linda C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 2007.

[7] Christian Bizer and Tom Heathand Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.

[8] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Jorunal of Web Semantics*, 7(1), 2009.

[9] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3), 2009.

[10] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *ACM International Conference on Management of Data (SIGMOD)*, 2008.

[11] Danah Boyd and Kate Crawford. Six Provocations for Big Data. *Social Science Research Network Working Paper Series*, 2011.

[12] Martin Brümmer, Ciro Baron, Ivan Ermilov, Markus Freudenberg, Dimitris Kontokostas, and Sebastian Hellmann. DataID: Towards Semantically Rich Metadata for Complex Datasets. In *10th International Conference on Semantic Systems*, 2014.

[13] Jeremy Debattista, Christoph Lange, and Sören Auer. daQ, an Ontology for Dataset Quality Information. In *7th International Workshop on Linked Data on the Web (LDOW)*, 2014.

[14] Bakshy Eytan, Rosenn Itamar, Marlow Cameron, and Adamic Lada. The role of social networks in information diffusion. In *21th International Conference on World Wide Web (WWW'12)*, 2012.

[15] Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. In *11th European Semantic Web Conference (ESWC)*, 2014.

[16] Annika Flemming. Quality Characteristics of Linked Data Publishing Datasources. Master's thesis, Humboldt-Universitt zu Berlin, 2010.

[17] Giorgos Flouris, Yannis Roussakis, and M Poveda-Villalón. Using provenance for quality assessment and repair in linked open data. In *2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn'12)*, 2012.

[18] Philipp Frischmuth, Sören Auer, Sebastian Tramp, Jörg Unbehauen, Kai Holzweißig, and Carl-Martin Marquardt. Towards Linked Data based Enterprise Information Integration. In *Workshop on Semantic Web Enterprise Adoption and Best Practice Co-located with 12th International Semantic Web Conference (ISWC'13)*, 2013.

[19] Philipp Frischmuth, Jakub Klímek, Sören Auer, Sebastian Tramp, Jörg Unbehauen, Kai Holzweißig, and Carl-Martin Marquardt. Linked Data in Enterprise Information Integration. *Semantic Web Journal*, 2012.

[20] Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing Linked Data Mappings Using Network Measures. In *9th European Semantic Web Conference (ESWC)*, 2012.

[21] Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayer. SCOVO: Using Statistics on the Web of Data. In *6th European Semantic Web Conference on The Semantic Web (ESWC)*, 2009.

[22] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In *3rd International Workshop on Linked Data on the Web (LDOW)*, 2010.

[23] Aidan Hogan, JüRgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics*, 2012.

[24] Renato Iannella and James McKinney. vCard Ontology - for describing People and Organizations. W3C Interest Group Note, 2014. http://www.w3.org/TR/vcard-rdf.

[25] Antoine Isaac and Ed Summers. SKOS Simple Knowledge Organization System Primer. W3C Working Group Note, 2009.

[26] Robert Isele, Jürgen Umbrich, Christian Bizer, and Andreas Harth. LDspider: An Open-source Crawling Framework for the Web of Linked Data. In *9th International Semantic Web Conference (ISWC), Posters & Demos Track*, 2010.

[27] Beverly K. Kahn, Diane M. Strong, and Richard Y. Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 2002.

[28] C. Maria Keet, María del Carmen Suárez-Figueroa, and María Poveda-Villalón. The Current Landscape of Pitfalls in Ontologies. In *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2013.

[29] Charles J. Kowalski. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society*, 1972.

[30] Sarasi Lalithsena, Prateek Hitzler, Amit Sheth, and P. Jain. Automatic Domain Identification for Linked Open Data. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 205–212, 2013.

[31] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S. Hopkins, and Nina Kruschwitz. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 2011.

[32] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-O: The PROV Ontology. W3C Recommendation, 2013. http://www.w3.org/TR/prov-o.

[33] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246, 2002.

[34] Joseph Juran. M. and A. Blanton Godfrey. *Juran's quality handbook*. McGraw Hill, 1999.

[35] Christian Mader, Bernhard Haslhofer, and Antoine Isaac. Finding quality issues in SKOS vocabularies. *Theory and Practice of Digital Libraries*, 2012.

[36] Nicolas Marie, Fabien Gandon, Myriam Ribière, and Florentin Rodio. Discovery Hub: On-the-fly Linked Data Exploratory Search. In *The 9th International Conference on Semantic Systems*, 2013.

[37] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems*, 2011.

[38] Nandana Mihindukulasooriya, Raul Garcia-Castro, and Miguel Esteban Gutiérrez. Linked Data Platform as a novel approach for Enterprise Application Integration. In *4th International Workshop on Consuming Linked Data (COLD'13)*, 2013.

[39] Tommaso Di Noia, Roberto Mirizzi, Vito Ostuni Claudio, Davide Romito, and Markus Zanker. Linked Open Data to Support Content-based Recommender Systems. In *8thInternational Conference on Semantic Systems - I-SEMANTICS '12*, 2012.

[40] Lawrence Page, Sergey Brin, Motwani Rajeev, and Winograd Terry. The PageRank Citation Ranking: Bringing Order to the Web, 1998.

[41] Eric Peukert, Julian Eberius, and Erhard Rahm. A Self-Configuring Schema Matching System. In *IEEE 28th International Conference on Data Engineering (ICDE'12)*, 2012.

[42] Emmanuel Pietriga, Christian Bizer, David Karger, and Ryan Lee. Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In *5th International Semantic Web Conference (ISWC'06)*, pages 158–171, 2006.

[43] Dave Reynolds. The Organization Ontology. W3C Recommendation, 2014. http://www.w3.org/TR/vocab-org.

[44] Massimiliano Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th International Semantic Web Conference (ISWC)*, 2014.

[45] Dagobert Soergel. Thesauri and ontologies in digital libraries. In *2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002.

[46] Umberto Straccia and Raphaël Troncy. oMAP: Combining Classifiers for Aligning Automatically OWL Ontologies. In *6th International Conference on Web Information Systems Engineering*, 2005.

[47] Fabian Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *16th International World Wide Web Conference (WWW'07)*, 2007.

[48] Osma Suominen and Eero Hyvönen. Improving the quality of SKOS vocabularies with skosify. In *The 18th International Conference on Knowledge Engineering and Knowledge Management*, 2012.

[49] Osma Suominen and Christian Mader. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, 2013.

[50] Holger Wache, Thomas Vögele, Ubbo Visser, Heiner Stuckenschmidt, Gerard Schuster, Heiko Neumann, and Sebastian Hübner. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *IJCAI Workshop: Ontologies and Information*, pages 108–117, 2001.

[51] Richard Y. Wang and Diane M. Strong. Beyond Accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 1996.

[52] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality Assessment Methodologies for Linked Open Data. *Semantic Web Journal*, 2012.