SENSE4US

Initial Investigation into Tools & Techniques for Semantic Linking & Consolidation of Heterogeneous Open Data

Project acronym:	**SENSE4US**
Project full title:	**Data Insights for Policy Makers and Citizens**
Grant agreement no.:	**611242**
Responsible:	**Timo Wandhöfer**
Contributors:	**Cristina Sarasua, Benjamin Zapilko, Leon Kastler**
Document Reference:	**D4.2.1**
Dissemination Level:	**PU**
Version:	**1**
Date:	**02/04/2015**

Project co-funded by the European Union under the Seventh Framework Programme

History

Version	*Date*	*Modification reason*	*Modified by*
0.1	20/03/2015	Initial draft	Timo Wandhöfer, Cristina Sarasua, Benjamin Zapilko, Leon Kastler
0.2	26/03/2015	Quality check	Somya Joshi
0.3	30/03/2015	Updated draft	Timo Wandhöfer, Cristina Sarasua, Benjamin Zapilko
1.0	02/04/2015	Final reviewed deliverable	Timo Wandhöfer

Table of contents

History **2**
Table of contents **3**
List of figures **5**
List of tables **6**
Glossary **7**
Executive summary **10**
Introduction **11**
1 Overview and Perspective **13**
1.1 Conceptual Overview 13
1.2 Perspective of Actors 14
1.2.1 Policy Maker 15
1.2.2 Project Partner 17
1.2.3 Data Publisher 18
1.3 Future Work 19
2 Data Selection within the Domain of Sense4us **20**
2.1 Real World Example 20
2.2 Potential Data Sources 24
2.3 Exemplary Data Publisher 27
3 Defining or reusing vocabularies **28**
4 Transformation of non-RDF to RDF data **29**
5 Data Interlinking on the Web of Data **31**
5.1 Existing approaches for knowledge integration on the Web of Data 32
5.1.1 The benefit of interlinking for Sense4Us end-users (policy makers) 33
5.1.2 Microtask crowdsourcing 34
5.1.3 Crowdsourced Interlinking 35
5.1.4 Preliminary experiments 36
5.1.4.1 Data sets and ground truth 36
5.1.5 Methodology 37
5.1.5.1 EventMedia and NYT toy experiments 38
5.1.5.2 Person11-Person12 experiment 38
5.1.6 Results 41
5.1.7 On-going and future work 43
6 Crowd Work CV **44**

6.1 Existing approaches for crowd profiling and task assignment 44
6.1.1 Recognition for Micro Work 45
6.2 Crowd Work CV ontology 45
6.2.1 Why an RDF-based data model? 47
6.2.2 Crowd Work CV data management 48
6.2.3 Crowd Work CV ontology verification 49
6.2.4 Future work 49
7 Publishing the interlinked Data 50
8 Linked Data validation and documentation 51
9 Conclusions 52
10 References 53

List of figures

Figure 1: Conceptual presentation of D4.2.1 13
Figure 2: Conceptual perspective for end users (policy maker) 15
Figure 3: Conceptual perspective for project partner 17
Figure 4: Conceptual perspective of data publisher 18
Figure 5: Conceptual perspective of the future work in WP4 19
Figure 6: Landing page Statistical Office for Berlin-Brandenburg 21
Figure 7: Interface to retrieve the database data of the Statistical Office for Berlin-Brandenburg 22
Figure 8: Statically data sheets hosted by the Statistical Office for Berlin-Brandenburg 23
Figure 9: Time serious in XSLX format hosted by the Statistical Office for Berlin-Brandenburg 23
Figure 10: Exemplary PDF representation on a HTML based website 27
Figure 11: RDF prototype – overview of resources 30
Figure 12: RDF prototype - single PDF resource 30
Figure 13: Examples of links on the Web of Data 31
Figure 14: Interlinking example 34
Figure 15: Overview of the process of crowdsourcing interlinking 35
Figure 16: Example RDF/XML notation 39
Figure 17: User interface 1 – Interlinking microtasks 40
Figure 18: User interface 2 - Interlinking microtasks 40
Figure 19: Judgements per contributor 42
Figure 20: Overview of the Crowd Work CV ontology, for describing agents, their user accounts, CVs, qualifications, work experiences, microtasks, their master microtasks, and marketplaces. 46
Figure 21: Crowd Work CV data management workflow 48

List of tables

Table 1: Glossary *9*
Table 2: Preliminary overview of suggested data sources *24*
Table 3: Suggested data sources (version 02/04/2015) *26*
Table 4: Accuracy of the crowd in crowdsourced interlinking – Input half correct links, half incorrect links *41*
Table 5: Accuracy of the crowd in crowd sourced interlinking – Input result of Silk *41*
Table 6: Experiment results *41*
Table 7: Lessens from experiments in online labour marketplaces *43*

Glossary

Term	*Explanation*
Crowd Worker	A person who accomplishes microtasks at online labour marketplaces.
Data Catalogue	Web repository containing a registry of data sets. Data catalogues are usually exposed via a Web portal and can also be accessed programmatically. They are useful for query information about the provenance, content and access information of data sets. One of the most outstanding data catalogues in Open Data is Datahub.io, published by CKAN. Every LOD cloud data set is registered in Datahub.io.
Data Publisher	A person or institution that publishes data. Usually data publishers own the data, but a data publisher may also be a technical organisation that aids data owners in making their data available (e.g. in the LOD cloud).
Data Server	A physical Web server that offers data. The data provider that owns the data, maintains this server.
Data Set or Dataset	As a general term, data set describes a certain amount of data bundled together and provided by a single data source. In RDF, a data set is a set of RDF statements. Example: the English DBpedia.
Data Source	A data source provides several data sets to others. Example: DBpedia.
DBpedia	"*DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web*." [see http://dbpedia.org/, retrieved on 18/03/2015]
End user	See policy maker
Entity	see Resource
Graph	A mathematical model for describing information. Consists of nodes and edges. The RDF model resembles as graph model.
Interlinking	In RDF, interlinking means to connect two different data sets via RDF links with each other.
Linked Open Data (LOD) Cloud	A collection of public available interlinked RDF data sets.

(Online Labour) Marketplace	Online crowdsourcing Web platform that offers microtasks published by requesters to be solved by crowd workers. Example: Amazon Mechanical Turk.
Microtask	A simple task that can be solved within short time by a person. Example: categorizing an image.
Microtask Crowdsourcing	Type of crowdsourcing characterized by outsourcing simple tasks to a potentially large group of people in the form of an open, in return of a small economic reward [Howe, 2006 http://archive.wired.com/wired/archive/14.06/crowds.html]
NRW	North Rhine-Westphalia (German state with the highest population)
Ontology	Explicit specification of conceptualization.[Gruber, Tom (1993); "A Translation Approach to Portable Ontology Specifications", in Knowledge Acquisition, 5: 199-199] Describes concepts existing in the data set and their relations between each other. In the given setting, the term is synonymous with the terms schema and vocabulary. Example: The DBpedia ontology (http://wiki.dbpedia.org/Ontology)
Policy Maker	An end user of the Sense4us toolkit that has direct or transferred responsibilities during a policy creation process. This can be for instance, a member of a parliament.
Resource Description Framework (RDF)	The Resource Description Framework (RDF) is a framework for expressing information about resources.[http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/ accessed 18.03.2015 14:30 CET)] RDF resembles a data model for semantic information. It consists of statements that relate a resource (called subject) with a value (called object) through a property (called predicate).
RDFS	RDF Schema provides a data-modelling vocabulary for RDF data. It is an extension of the basic RDF vocabulary. [http://www.w3.org/TR/2014/REC-rdf-schema-20140225/ (accessed 18.03.2015 14:30 CET)]
Requester	A person or institution that offers microtasks in online labour marketplaces in exchange of money.
Resource	Describes something in the world, like a person, a Web source, or an abstract concept. In the Semantic Web / Linked Data scenario, resources are described with RDF statements.
Schema	see Ontology

Vocabulary	See Ontology
Work package (WP)	Work package in combination with a number refers to one of eight work packages of the Sense4us project, which are mentioned in the description of work (DoW).

Table 1: Glossary

Executive summary

The work we are presenting in this deliverable is twofold:

On the one hand, we aim at extending the Linked Open Data (LOD) cloud with datasets referring specifically to the domain of Sense4Us. While the LOD cloud contains plenty of interesting datasets, we observe a lack of data about renewable energy. Hence, in order to enhance the pilots in which policy makers browse related information; we identify relevant open data that is available on the Web in a non-RDF format, and publish it scenario based as Linked Data, by processing the five steps *Defining or reusing vocabularies*, *Transformation of non-RDF to RDF data*, *Data Interlinking on the Web of Data*, *Publishing the interlinked Data*, and finally *Linked Data validation and documentation*.

On the other hand, we investigate how to improve (semi-) automatic knowledge integration techniques, which are necessary to interlink datasets on the Web of Data. Our approach proposes the use of microtask crowdsourcing for extending (semi-) automatic interlinking techniques with human computation.

With these two major contributions, we target public organizations and private agents related to the domain in Sense4Us, who may be interested in publishing Linked Data. If data publishers follow our strategy, the visibility of their data could increase via those tools that are connected to the LOD cloud. One of these tools for instance is the Sense4us toolkit that will provide WP4 ranking mechanisms for LOD. Hence a Sense4us end user, the policy maker, could get richer results.

Before we state the goals for data interlinking within this deliverable, we provide a conceptual overview of the general work in WP4. Within this context we highlight the different views of the main actors: the data publisher, the crowd worker and the policy maker. And we provide a scenario-based introduction into data sources.

Introduction

Semantic Web technologies[1] [Antoniou et al., 2004] aim at enabling a Web of Data, in which distributed and heterogeneous datasets curated and maintained by different parties may be queried in an integrated way. Such integrated global data space is enabled by machine-readable descriptions of the data, which are created by means of Semantic Web standards like the Resource Description Framework (RDF) data model, and RDFS and OWL for defining ontologies (or vocabularies). By adding an RDF-based metadata layer to Web content and providing a separate and explicit definition of the semantics of the data, software applications may consume information spread over the Web and stored in formerly unknown data sources (e.g. relational databases, plain CSV files, text documents, HTML Web sites, etc.) in a unified manner [Heath et al., 2011].

With the improvement in technology and the emergence of new initiatives aiming to boost the creation of the Web of Data, the amount of semantic data published on the Web has increased considerably in the last years. One of the most remarkable efforts is the Linking Open Data community project[2], which developed several tools and defined best practises for various steps of the semantic data lifecycle. More specifically, the project focused on creating, integrating, publishing, documenting, and validating so-called Linked Data (i.e. data that follows the Linked Data principles)[3]. As a result, the Linked Open Data[4] cloud was created. The LOD cloud is a set of RDF datasets interlinked with each other, containing as of August 2014 datasets about several topical domains such as media, life sciences, government, publications, linguistic resources and social networking [Schachtenberg et al., 2014].

This deliverable covers the initial investigations of WP4 into "Tools & Techniques for Semantic Linking & Consolidation of Heterogeneous Open Data". The standard process for Linked Open Data publication includes the five steps of *Defining or reusing vocabularies*, *Transformation of non-RDF to RDF data*, *Data Interlinking on the Web of Data*, *Publishing the interlinked Data*, and finally *Linked Data validation and documentation*. Additionally to this process the deliverable provides a conceptual overview of how each of the five steps fit together, how the whole process fit into the Sense4us project, and finally who the beneficiary are. Hence we collected a list with Open Data sources that is related for the Sense4us scenario renewable energy. To address these issues the structure of this deliverable is as following:

In section 1 we provide a conceptual overview how the "five steps" fit together and what the roles of WP4 members University of Koblenz and GESIS are for developing an integral strategy. Then proceed with different perspectives by introducing different actors and how they influence or benefit from the Linked Open Data publication process.

In section 2 we take a look at Open Data, which is not available in RDF format or interlinked to the cloud. Here we start with a real life data source example, mentioned by a member of the Berlin City Parliament during the first phase of end user engagement. The project partners have contributed to a compilation of data sources that can be scenario-based used during the

[1] URL: http://www.w3.org/standards/semanticweb/ (Retrieved on 10/03/2015).

[2] URL: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData (Retrieved on 10/03/2015).

[3] URL: http://www.w3.org/DesignIssues/LinkedData.html (Retrieved on 10/03/2015).

[4] URL: http://lod-cloud.net/ (Retrieved on 10/03/2015).

upcoming pilots. After the overview of potential data sources we introduce a typically governmental body, who publishes official Open Data within the format of PDF documents.

In section 3 we pick up a governmental body as exemplary data publisher. Within the **first step *Defining or reusing vocabularies*** is important to define the metadata that is going to be used to describe the data. A good practice in ontology engineering is to reuse existing and well-adopted ontologies. The LOV[5] Web portal provides a repository of Linked Open Vocabularies that may be reused either partially or totally.

In section 4 we exemplary show the **second step *Transformation of non-RDF to RDF data***. If the data publisher owns data that is in a non-RDF format, then this step is known as RDFizing the data. This step usually entails the use of a mapping tool that defines the way in which for example the data of a relational database needs to be translated into RDF data.

In Section 5 we introduce the reader to the **third step for *Data Interlinking on the Web of Data***. We explain the way interlinking benefits the end-user of the Sense4Us scenario and describe our research on using microtask crowdsourcing in a hybrid-interlinking context (i.e. combining machine and human computation).

The challenges that we faced during this research on the application of microtask crowdsourcing in a knowledge integration scenario motivated the contribution we present in section 6. The *Crowd Work CV* (Curriculum Vitae) is a RDF-based data model that aims to improve the current landscape of crowdsourcing marketplaces in terms of recognition for work.

In section 7 we explain the **fourth step *Publishing the interlinked Data*** in order to enable dereferenceable URIs (i.e. URIs that provides RDF descriptions in response to HTTP GET requests).

In section 8 we show the **fifth and last step *Linked Data validation and documentation*** to satisfy all the Linked Data principles.

In section 9 we conclude with a summary of key insights emerging from this deliverable, and suggested next steps.

[5] URL: http://lov.okfn.org/dataset/lov/ (Retrieved on 10/03/2015).

1 Overview and Perspective

The aim of this section is to show how the WP4 work of University of Koblenz (LOD-Ranking & Data Interlinking) and GESIS (RDFization & Open Data) fit all together and can be beneficial for the identified actors.

1.1 Conceptual Overview

The following Figure 1 visualizes the aim of the task T4.2 conceptually and hence what this deliverable is mainly about. The figure includes four different kinds of elements (see legend, bottom line). The cloud icon (e.g. labelled "Linked Open Data Cloud", see right) represents an external data source. Data sources can be queried and provide input data for WP4 processes, which are displayed as rectangles. For instance each of the five steps mentioned above are representing a process. They are chained together and need an input and create an output. For example the process "Transformation of non-RDF to RDF Data" creates the artefact "RDF Data" (displayed as rectangle with round corners). Together all elements that are part of the middle section compose the whole process for "Linked Open Data Publication" (big orange coloured box, see background). This main process "Linked Open Data Publication" needs the input of "Non-RDF Open Data" (see left) and "Linked Open Vocabulary" (see left). The "Crowd" is the only actor within the following figure (actors are displayed as an eclipse icon, see left). The "Crowd" takes the action for interlinking and influences therefore the links between the LOD Cloud and RDF Open Data sources. The "Crowd" is one of four actors we have identified within WP4 in total. The perspective of "Policy Maker", "Project Partner" and "Data Publisher" will be introduced in the following section 1.2 individually.

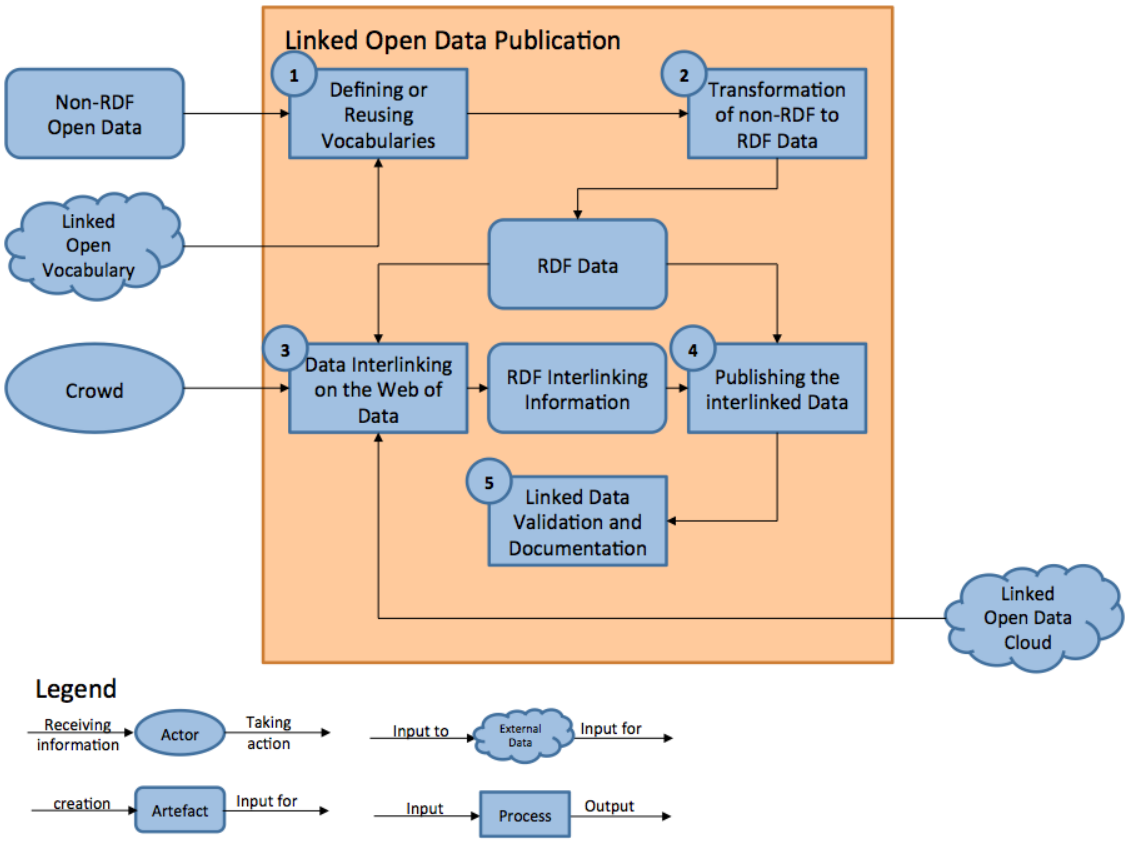

Figure 1: Conceptual presentation of D4.2.1

1.2 Perspective of Actors

The aim of the Sense4us project is to use state of the art information and communication technologies (ICT) to support the policy decision-making process on the EU level, the national level and the state/local level. The focus is on Open Data (Cp. WP4), Social Networks with Twitter in particular (Cp. WP5) and policy simulation (Cp. WP6). The end user can use the tools via the Sense4us toolkit, which comes with a user interface (Cp. WP3). The main benefit of the toolkit is that the end user engages with different kinds of technical modules without the necessity of having a technical background. Hence the challenge is not only creating an easy to understand user interface; the challenge is the "transparency" how the tools work and the "trustfulness" that different actors using Sense4us within everyday life.

The following sections explaining three distinguish actors:

- The policy maker's aim is to make sense of the search, analysis and simulation tools for using the results within the policy making process. Hence the user interface is the most important part for the policy maker.
- The project partners are using the tools' functionalities' via an application-programming interface (API) instead of the user interface.
- The data publisher is an important actor for enriching the LOD-Ranking. They are not using the toolkit's user interface or its APIs. Data publishers get recommendations on strategies, how to interlink their data with the LOD Cloud.

The following figures (Figure 2, Figure 3, Figure 4) include similar elements as Figure 1. To keep the figures simple we did not include the legend (Cp. Figure 1) again. Every figure includes all of the three actors to see which role they fit within the work package. But for a better understanding there is only one actor activated per figure and the other actors are greyed out.

1.2.1 Policy Maker

Figure 2 displays the conceptual view how the "Policy Maker" might benefit from the work done in WP4. The policy maker (see left) interacts with the Sense4us user interface (see green box in the background called "User Interface"). He has two options for retrieving the Linked Open Data Cloud. The outcome of both retrieval strategies is a ranked list of data sets. Imagine the policy maker needs to know *the different kinds of renewable energy suppliers and their proportion on the general energy production in the German state North Rhine-Westphalia*.

Scenario 1: The policy maker has a technical report addressing the topic *renewable energy in the German state North Rhine-Westphalia*. Hence he is interested to know the relevant data sets concerning the document's content. Therefore the policy maker uploads the PDF document (see "Policy Document (Draft)") and receives a ranked list of data sets (see "Data Set Ranking") and the structure of their connectivity (see "Connectivity Structure"). The process behind this output is the automatically analyse concerning the occurrence of words within the PDF document (see "Topic Model Analysis"). Often mentioned words that are connected frequently will be clustered to a set of words (here "Topics"). These topics will be modified (see "Entity Resolution") to run proper searches on the LOD cloud and its connected Open Data sources. How this works in detail is on-going work (Cp. section 1.3) and is scheduled for the next deliverable. This approach can be done with nearly every text-based PDF document (e.g. a draft bill, a petition, a Hansard or a newsletter).

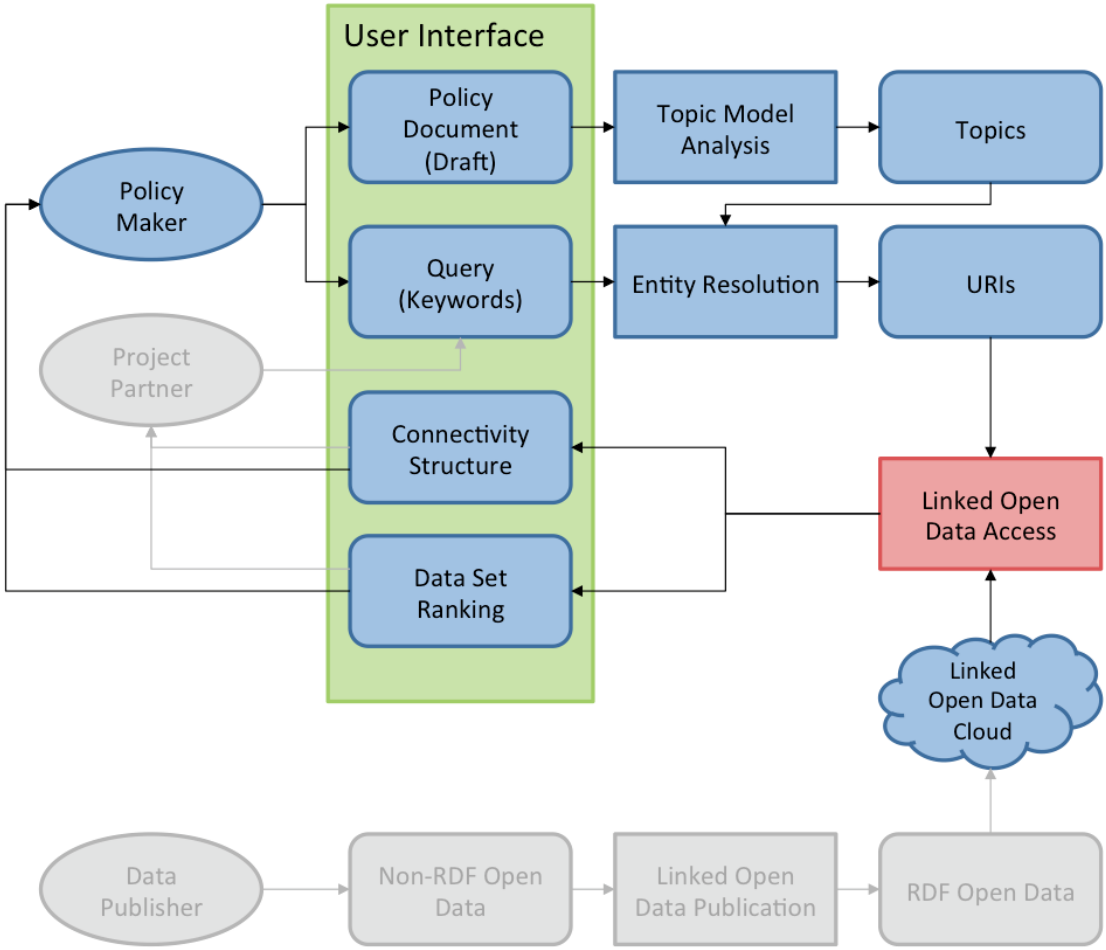

Figure 2: Conceptual perspective for end users (policy maker)

Scenario 2: The policy maker has the keywords *renewable energy*, *North Rhine-Westphalia* and *proportion* in mind. Hence he starts a query via the Sense4us dashboard (see "Query (Keywords)") and receives a ranked list of data sets (see "Data Set Ranking") and the structure of their connectivity (see "Connectivity Structure"). These search words will be modified (see "Entity Resolution") to run proper searches on the LOD cloud and its connected Open Data sources. How this works in detail is on-going work (Cp. section 1.3) and is scheduled for the next deliverable. This approach can be done with nearly every keyword as search term.

How to address the upcoming pilots: As mentioned above the scalability of Linked Open Data Publication is on providing a strategy how data publisher can interlink their data to the LOD cloud. Hence our approach is a scenario-based implementation of at least one strategy. The criteria for choosing one scenario refer to its usefulness for end users who will be engaged in the validation trails. Apart from this the chosen scenarios should be in the Sense4us topic area – which is renewable energy. Third the scenario should be "representative" regarding further potential data sets. The aim of the scenario-based approach is its scalability for data publishers.

1.2.2 Project Partner

In addition to the policy maker who uses the tool's component via the Sense4us user interface, project partners can use the WP4 components additionally via an API. Project partners can use the outcome of the "Data Set Ranking" or also provide keywords for starting a query, which result from previous analysis.

Scenario: A policy maker uploads a technical report that summarizes the different kinds of technologies for renewable energy, which are implemented in the German state of North Rhine-Westphalia. With the Sense4us tool "policy model assist" (integrated in the Sense4us prototype) the policy maker for instance gets a list of "concepts" that are included in the document: e.g. *wind power*, *solar energy* or *biofuel energy*, *state North Rhine-Westphalia*. Given this output these keywords could be used for starting the WP4 process of "Data Set Ranking". The policy maker may receive relevant information from the "Linked Open Data Cloud" that e.g. summarizes the proportion of energy production for each of the renewable energy technologies concerning the whole production of energy production of North Rhine-Westphalia. This data could be interlinked by a local data publisher within a previous step by implementing the WP4 "Linked Open Data Publication" strategy (cp. Figure 4 in addition).

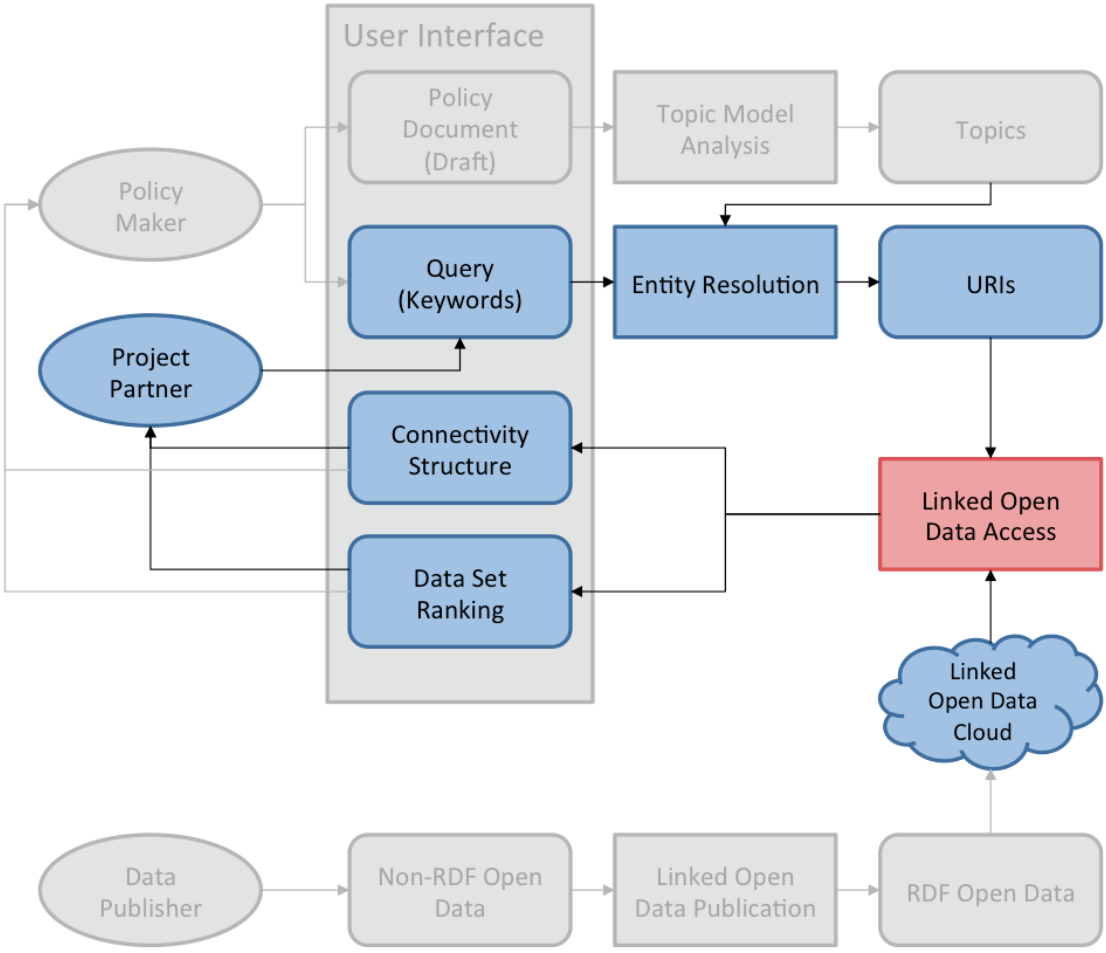

Figure 3: Conceptual perspective for project partner

1.2.3 Data Publisher

The data publisher is the central actor for the work within this deliverable. The more and the better the data publishers interlink their data with the LOD cloud the richer will be the results for LOD data sets for the policy makers. This again refers to the five steps mentioned within the introduction. Hence the scalability of WP4 is not processing as much data sources as project resources are available; the scalability of WP4 results in recommendations for strategies, how potentially all data publishers with relevant data interlinking their data.

Figure 4 displays the conceptual view for data publisher (see left). It is not always the case that the Open Data is available in the RDF format. Hence the whole process "Linked Open Data Publication" is recommended (Cp. Figure 1). If the data is already in a proper RDF format with a proper vocabulary the first two steps can skipped. Figure 4 shows as benefit for the data publisher a new connection for accessing the Open Data via Linked Open Data tools like Sense4us. More details regarding the five steps can be found in sections

- Step 1: Cp. section **3 Defining or reusing vocabularies**;
- Step 2: Cp. section **4 Transformation of non-RDF to RDF data**;
- Step 3: Cp. section **5 Data Interlinking on the Web of Data** and **6 Crowd Work CV**;
- Step 4: Cp. section **7 Publishing the interlinked Data**;
- Step 5: Cp. section **8 Linked Data validation and documentation**.

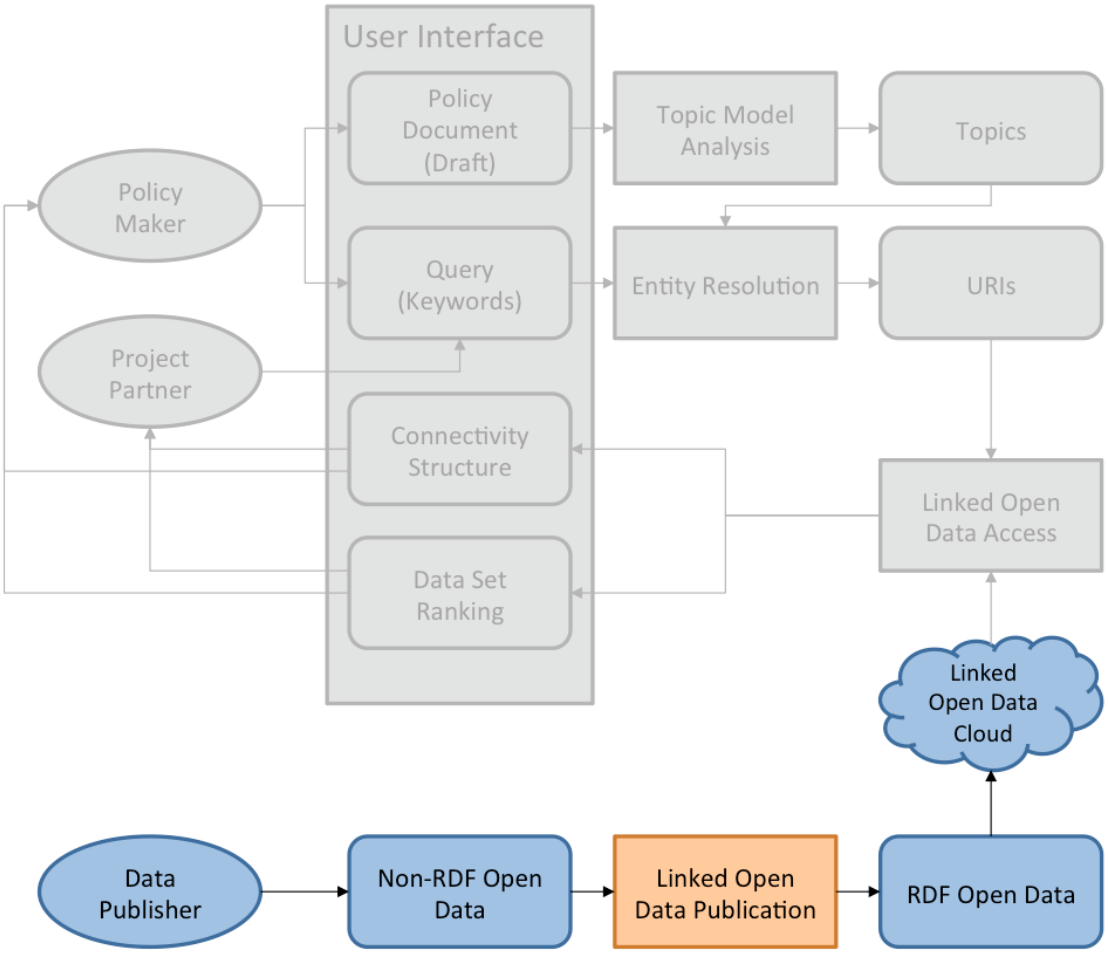

Figure 4: Conceptual perspective of data publisher

1.3 Future Work

A central task of WP4 in the future is the improved accessibility of the Linked Open Data cloud for policy makers as well as project partners. The two major elements here are the ranking of available data sets within the Linked Open Data cloud, in regard to given queries and related URIs, and the discovery of relevant connections between these URIs (Cp. Figure 5).

Both elements support policy makers to gain deeper insight in topics as well as providing them with additional information about related fields and possible effects of a policy. These elements are fields of research for us and will be addressed in detail within the deliverables D4.1.2 and D4.3.2.

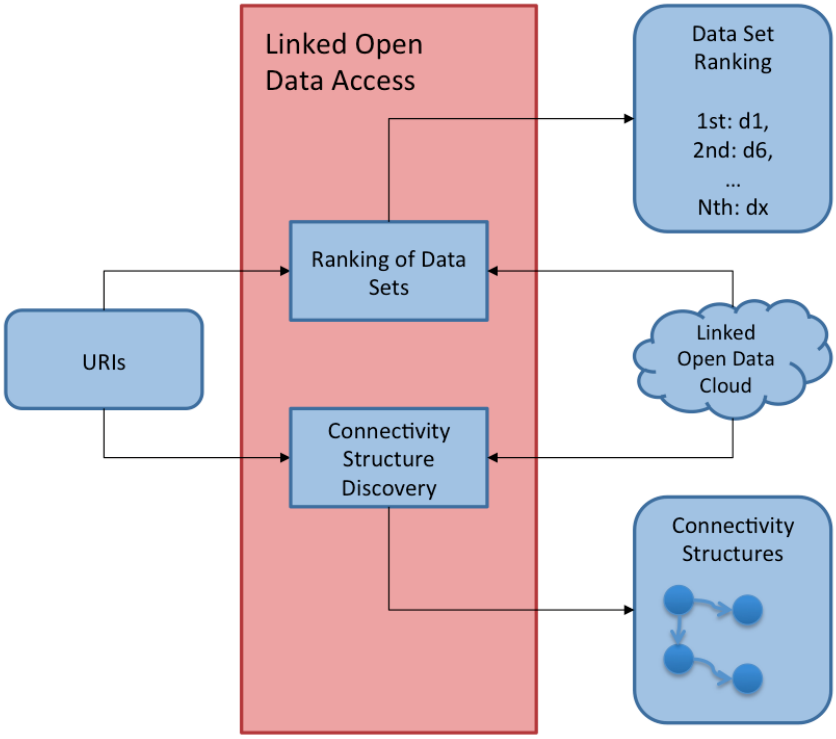

Figure 5: Conceptual perspective of the future work in WP4

2 Data Selection within the Domain of Sense4us

2.1 Real World Example

Beside the data available via the LOD cloud there exist a vast amount of Open Data, which is not available in RDF format or interlinked to the cloud. For instance one MP from the Berlin City Parliament mentioned during the first phase of end user engagement using the local database "Statistik Berlin Brandenburg" for the decision making process (see deliverable D2.1, p.71). Responsible for its content is the "Amt für Statistik Berlin-Brandenburg" (Statistical Office for Berlin-Brandenburg), who is "*the central service provider for statistics concerning the states Berlin and Brandenburg. The institute performs services in the area of information and data analyses for the public, the public administration and politics as well as for science and industry.*

Core business of the institute is the production of official statistics for Berlin and Brandenburg. The institute collects, processes, analyses and interprets data and publishes the results. The basic statistics are published on internet and can be requested free of charge at the Information Management. Additionally the institute offers special standard reports for fixed prices. Furthermore users can request special analyses which will be charged on a time and material basis."[6]

Within this section we will have a closer look at this real world data publisher to see what kind of information are available and what the different formats look like and how the data can be retrieved.

[6] Cp. URL: https://www.statistik-berlin-brandenburg.de/inhalt/inhalt-impressum.asp (Retrieved on 10/03/2015).

Figure 6 shows a screenshot of the website's landing page. On the left hand side the user can navigate through the menu of different kinds of provided information: e.g. job offerings, newsletter, events, projects, reports and publications. Current press releases and the current topic can be found in the middle part. In contrast to the reports and publications, which are generally available as PDF or XLS format, the current topic (here "Tourism in Berlin Brandenburg 2014") is visualized as an interactive map. The bigger map on the right hand side represents the state Brandenburg including its regions. The city state Berlin is directly in the middle of Brandenburg. On the map Berlin is displayed in a bigger scale (see left), because its divisions are smaller in contrast to Brandenburg. Hence the map representing Berlin's divisions is displayed on the left hand side. The aim of the map is the visualization of 'numbers for overnight stays within the year 2014'. For instance in the Berlin centre division 34036,2 overnight stays per 1000 habitants where registered.

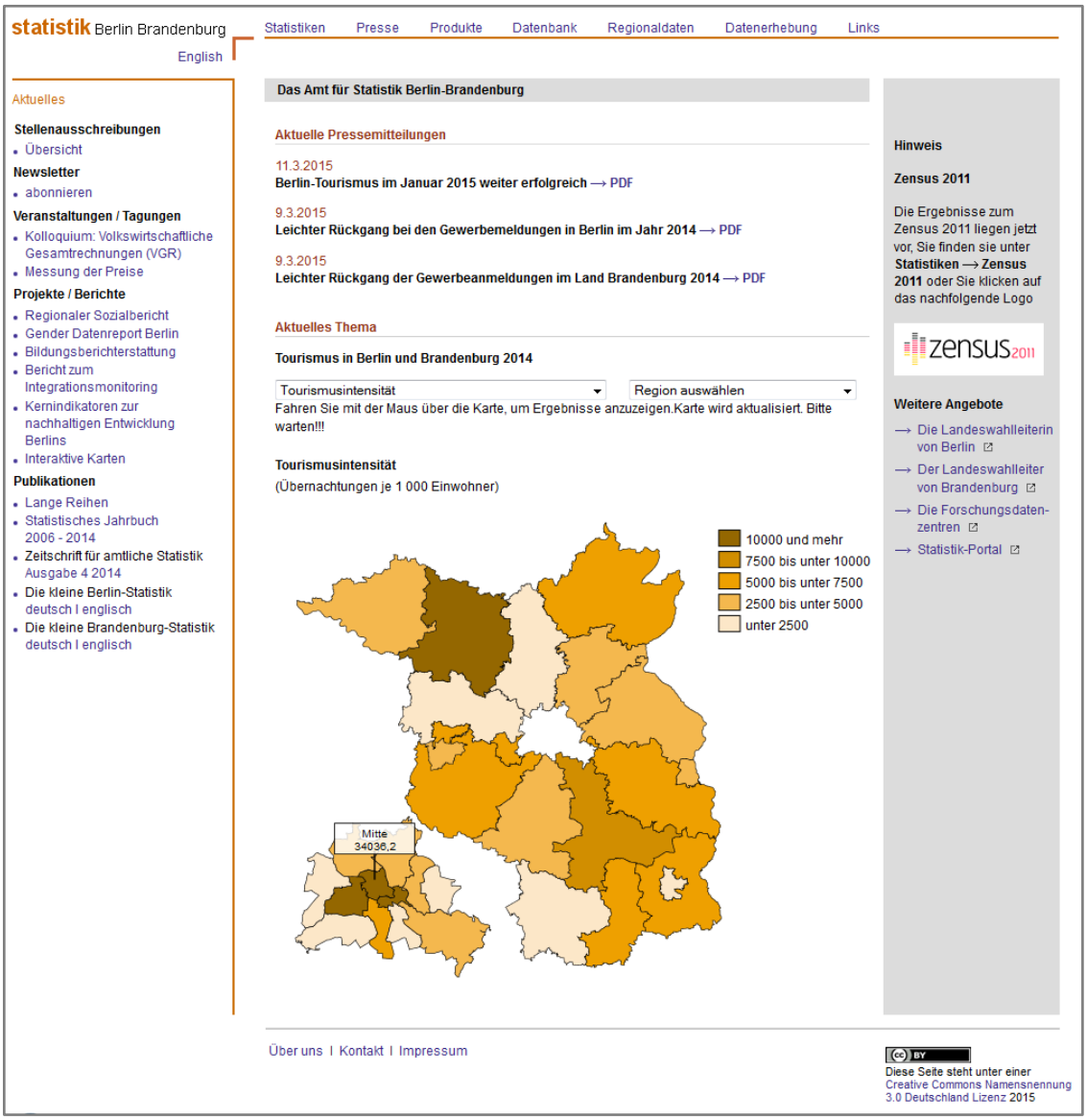

Figure 6: Landing page Statistical Office for Berlin-Brandenburg[7]

[7] URL: https://www.statistik-berlin-brandenburg.de/ (Retrieved on 12/03/2015).

The data for overnight stays for the states Brandenburg and Berlin can also retrieved via a database interface. Figure 7 shows the interface with different kinds of filters (see left) and the result table that represent the parameters (see right).

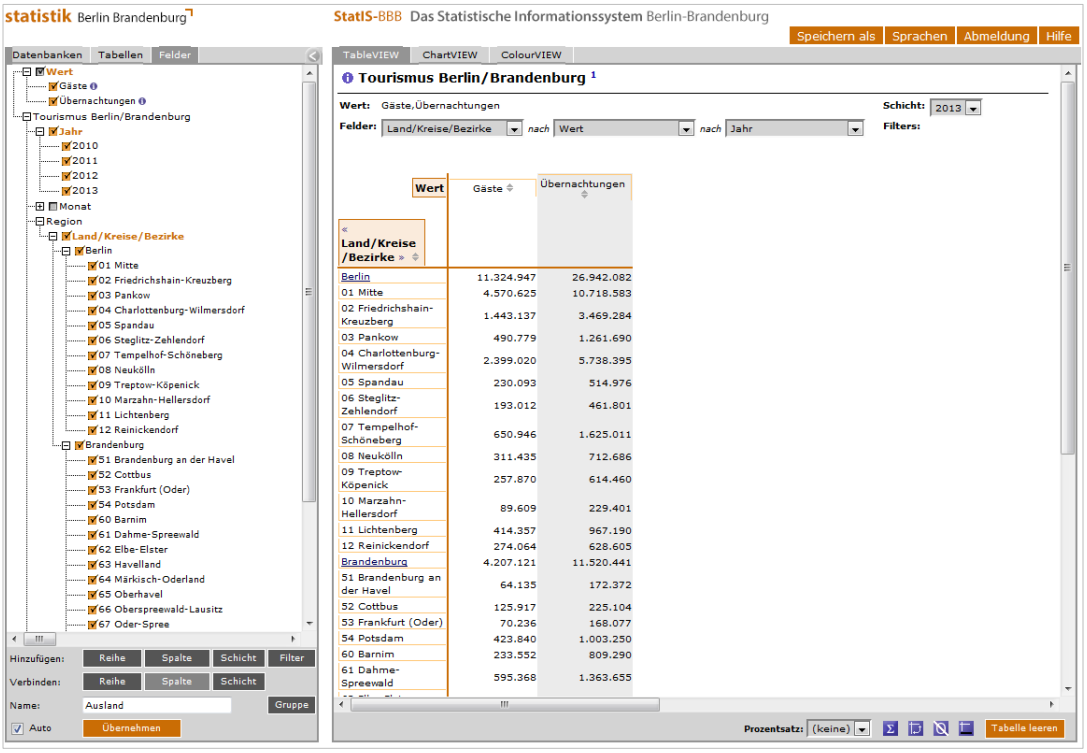

Figure 7: Interface to retrieve the database data of the Statistical Office for Berlin-Brandenburg

Additionally to the tables, which can be used interactively by selecting or deselecting filters, statically tables can be retrieved.

Figure 8 shows the 'spending on environment protection and individual products for the year 2011'.

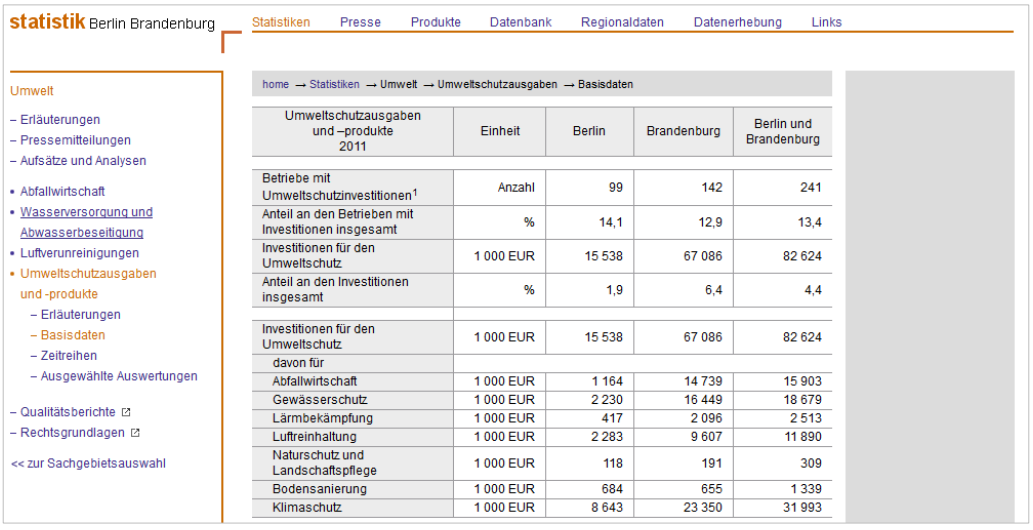

Figure 8: Statically data sheets hosted by the Statistical Office for Berlin-Brandenburg[8]

For instance in the category *Lange Reihen* (engl. time serious) the data is also available as XLSX format, which is an Excel Microsoft format for sheets.

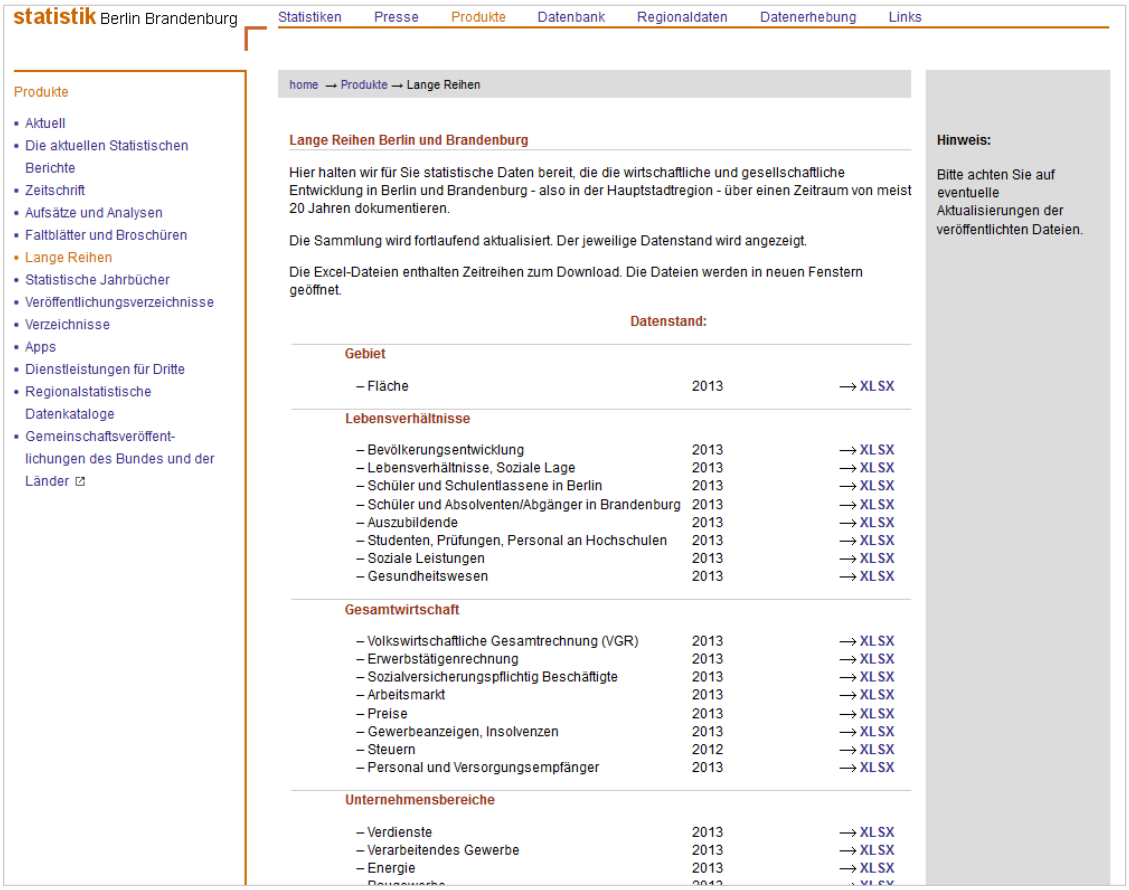

Figure 9: Time serious in XSLX format hosted by the Statistical Office for Berlin-Brandenburg[9]

There is more data, hosted by the Statistical Office for Berlin-Brandenburg available. But only these distinguished examples showcase the fact that Open Data respectively the way in which to retrieve it can get really complex to make sense of it. And there are more data publishers available.

[8] Cp. URL: https://www.statistik-berlin-brandenburg.de/BasisZeitreiheGrafik/Bas-Umweltschutzausgaben.asp?Ptyp=300&Sageb=32005&creg=BBB&anzwer=7 (Retrieved on 19/03/2015).

[9] Cp. URL: https://www.statistik-berlin-brandenburg.de/produkte/produkte-langereihen.asp (Retrieved on 19/03/2015).

2.2 Potential Data Sources

In general, open data sources are collected in portals on the Internet. But sources can also be an index of PDF documents or deep links that are available on a website, e.g. of a governmental agency. The benefit of open data portals is the huge amount of data. For instance, they may include 'smaller' databases or indexes, why the focus is on portals.

The Sense4us consortium suggested the following open data sources (Cp. Table 2), which are thematically related to the Sense4us scenarios (Cp.WP2 deliverables D2.1 & D2.3). The ranking of the list is the sequence when they were suggested.

Number	*Data source*
1	http://www.stat.io/about/product/
2	http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/
3	http://www.it.nrw.de/
4	http://publicdata.eu/
5	http://open-data.europa.eu/
6	http://www.offenedaten.de
7	https://www.govdata.de/
8	http://daten.berlin.de/
9	http://www.offenedaten-koeln.de/
10	http://data.un.org/
11	https://www.umwelt.nrw.de/english/
12	http://linkedpolitics.ops.few.vu.nl/
13	http://www.ons.gov.uk/ons/index.html
14	http://data.gov.uk/
15	http://www.legislation.gov.uk/
16	http://ec.europa.eu/public_opinion/index_en.htm

Table 2: Preliminary overview of suggested data sources

The suggested data sources in Table 2 were analysed against different categories. Table 3 shows an overview of the analysed categories (see first column) and sub categories (see second column). The third column showcases a particular data source. Here the selection was the *Ministry for Climate Protection, Environment, Agriculture, Nature Conservation and Consumer Protection of the German State of North Rhine-Westphalia*, who regularly publish

PDF documents concerning environmental topics relevant for North Rhine-Westphalia. The reason why this data source was selected is the compilation of technical reports concerning *renewable energy in the German state North Rhine-Westphalia*, which are relevant for policy makers in North Rhine-Westphalia, who are responsible for this topic. Similar to this ministry's website there are lots of data sources available. Hence we are using this governmental body as a representative data publisher with the characteristic "website providing (official) PDF documents".

Generally the summarization and categorization of data sources regarding Table 3 is on-going process. We frequently update the table regarding data sources. In addition we will formulate more characteristics for data publishers (e.g. Open Data portal with API or without API). The aim is to provide another scenario-based example for the RDFization process. The D4.2.2 deliverable (month 30) will provide a more improved version with a list of all of the proposed data sources and a status report concerning their analysis and data process.

The first category, displayed in Table 3, provides an overview of the relevant **metadata** of the data source: The **name** can be the name of the data source or even the provider of the data source. The **URL** links to the website with the general description and the data or a subset of the data. A **description** shows the reader what the data source is about. Policies generally refer to a special **area (geographically)** what the policy is about (here the state North Rhine-Westphalia in Germany). The **language** is important to provide data within the language the policy maker is possible to understand. Thematically the range of data of a potential source can be much brought. Therefore the **potential policy issues/topics** set limits for which fields the data is relevant. Furthermore, the metadata provides **contact details** and the information who **suggested** the data source.

Because the Sense4us project considers end-users in three different levels the category **application** shows even if the data is relevant for the **EU level**, the **national level (GB)** or the **national/state level (DE)**. From the technical point of view it is relevant which tool can be used to retrieve/analyse the data. The main use case is that the data will be accessed by the **Linked Open Data Access** (Cp. section 1.1 and 1.3) and can possibly processed by components in WP3 (e.g. topic analysis), WP5 (e.g. sentiment analysis) or WP6 (e.g. simulation).

When a potentially relevant data source was identified, its technical details are of high relevance. If an **API is available**, the data can be automatically accessed. A data source is **linked** when it contains links to other data sources. If an API exists, there are **several opportunities how the data can be retrieved automatically**. The data can have different **formats** (e.g. numeric data that is stored within a database or a table and PDF documents that are available by an index or deep links). We distinguish between **static or dynamical** updates of data sources. That means whether the offered data is dynamically updated together with its origin or whether it is updated by a particular update process that is conducted once in a particular time. This is directly connected with the information on **how frequently the data is updated**, which guarantees that the data is up to data. Apart from that it is relevant for the end-users when the **data were published or updated the last time**.

For the project and even the users of the data it is important to know the valid data policy. For instance **if the data is personal data**, is the **data public or private**, is the use of data **free of charge** and are there any **terms and conditions** that prevent users to use the data?

The last row shows the **status** how far the data source is analysed or processed. The example data source were successful analysed and is available for the Senese4us components via RDF.

Category	Subcategory	Data source 11
Metadata	**Name**	Ministry for Climate Protection, Environment, Agriculture, Nature Conservation and Consumer Protection of the German State of North Rhine-Westphalia
	URL	Website: https://www.umwelt.nrw.de/english/ (Retrieved on 19/03/2015).
	Description	*"The Department is becoming ever more active in supporting activities and financial assisting private individuals, community institutions and business enterprises. Today, the Ministry for Climate Protection, Environment, Agriculture, Conservation and Consumer Protection provides compliance assistance and initiates and supports volunteer activities in North Rhine-Westphalia, for example in the areas of eco-efficiency, conservation and consumer advice*." [https://www.umwelt.nrw.de/english/, retrieved on 19/03/2015]
	Area (geographically)	German State of North Rhine-Westphalia (abbreviated to NRW)
	Language	German
	Potential policy issues/topics	Renewable energy in NRW; Energy specifications for NRW
	Contact details	Imprint: https://www.umwelt.nrw.de/impressum/ (Retrieved on 19/03/2015).
	Suggested by	GESIS
Application	**EU level**	NO
	National level (GB)	NO
	National/state level (DE)	YES (NRW)
	Semantic search (WP4)	YES
	Sentiment (WP5)	NO
	Simulation (WP6)	MAYBE
Technical description	**API available?**	NO
	Is the data linked?	NO
	How we can access the data programmatically?	Only manually
	What formats is the data in?	Index of PDF documents
	Is the data static or dynamically updated?	Static updates
	What is the frequency of updates?	Looks like there are no updates. New documents will be added to the list.
	Date of the last update?	No updates available; Last document added July 2014.
Data policy	**Is the data personal data?**	NO
	Is the data public or private?	PUBLIC
	Any fees? Or is the data free of charge?	FREE
	Terms and conditions for accessing the data?	Not allowed to be used for campaign purposes.
Sense4us integration	**Status**	Available for semantic search (month twelve prototype)

Table 3: Suggested data sources (version 02/04/2015)

2.3 Exemplary Data Publisher

For illustrating the step Transformation of non-RDF to RDF data (Cp. section 4) we decided to use the data publisher *Ministry for Climate Protection, Environment, Agriculture, Nature Conservation and Consumer Protection of the German State of North Rhine-Westphalia.* The ministry is part of the government of the German state of North Rhine-Westphalia and is therefore engaged in the policy making process. The work of initial end user engagement of WP2 has identified that governmental bodies, which are publishing official data, are relevant for policy makers (Cp. Deliverable D2.1, p.72f.). There are lots of data sources available similar to the ministry's website. Hence we are using this governmental body as a representative data publisher with the characteristic "website providing (official) PDF documents", because the data source is part of data compilation (Cp. Table 3). Figure 10 an exemplary PDF representation on a HTML website. The technical report covers facts and figures concerning energy data in NRW for the year 2013.

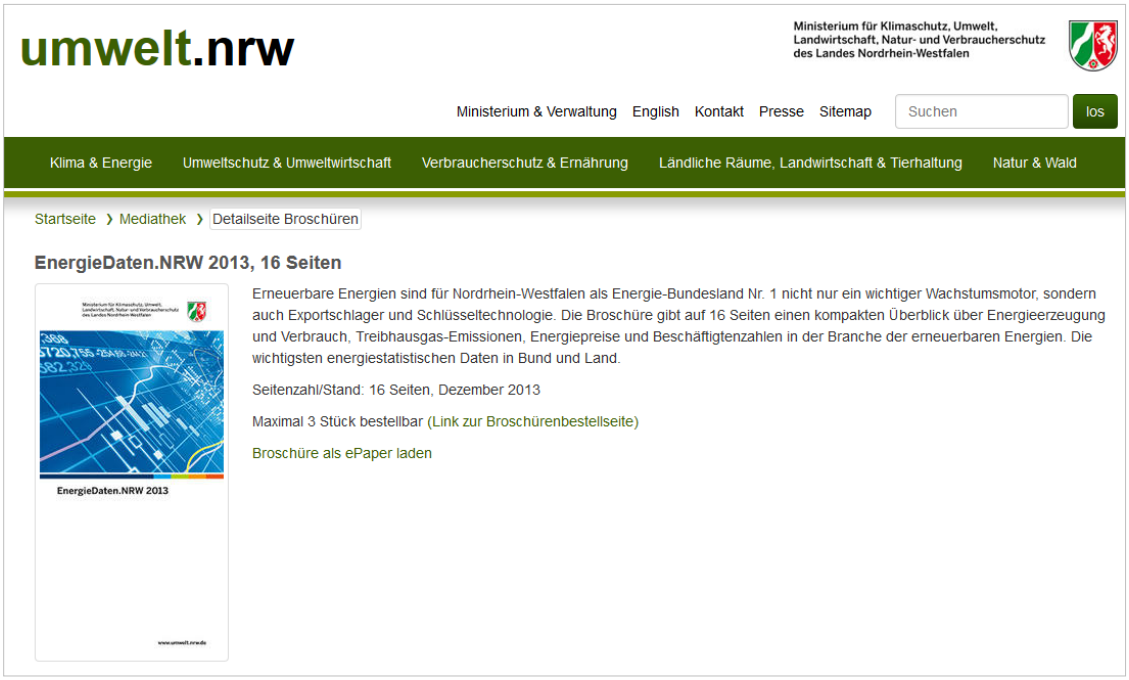

Figure 10: Exemplary PDF representation on a HTML based website[10]

[10] URL: https://www.umwelt.nrw.de/mediathek/broschueren/detailseite-broschueren/?broschueren_id=1408&backId=147&cHash=4ef1b10e9758348d7580ce50c30ba6ec (Retrieved on 04/03/2015).

3 Defining or reusing vocabularies

A benefit of representing data in RDF is that the semantics of the data, i.e. the meaning of particular entities, can be expressed and processed by machines. This is enabled by using particular vocabularies for describing single entities in the data, e.g. to express that a person is an author or that a specific resource is a journal article. Thus, the vocabulary design of data, which is intended to be represented in RDF, is a necessary step that has to be conducted before the RDFization of the data.

Various vocabularies are available for expressing particular types of data (e.g. publications, statistical data) or particular processes and activities in context of data (e.g. research processes, data documentation). In the Semantic Web community, it is seen as best practice to reuse existing vocabularies as much it is possible for ensuring interoperability with other data sets [Heath et al., 2011].

Revisiting the example of the *Ministry for Climate Protection, Environment, Agriculture, Nature Conservation and Consumer Protection of the German State of North Rhine-Westphalia* introduced in section 2.3, it is necessary to know what type of data is to be RDFized. Based on this information, the vocabulary design can be done, i.e. it can be determined for each element of the data by which vocabulary it should be expressed. In the example, there is a list of PDF documents, which are described by little metadata information like a short abstract, the publication year, etc. The DCMI Metadata Terms[11] are used for representing data about publications widely across disciplines and communities. Thus, it was decided to use terms of this vocabulary for expressing the publication data of the ministry.

For data publishers, it is necessary to determine in which extent and by which vocabularies their data should be expressed in RDF. The choice of vocabularies is relevant for ensuring interoperability with other RDF data sets. The extent of the data represented in RDF may influence the interlinking process, i.e. the more information is available about a specific resource like abstract, keywords, subjects, etc. the interlinking tool may provide better or more links to other data sets.

[11] Cp. URL: http://dublincore.org/documents/dcmi-terms/ (Retrieved on 11/03/2015).

4 Transformation of non-RDF to RDF data

The technical transformation of non-RDF data to RDF data can be conducted after the vocabulary design is finished. The expected output for the general approach was prototypical implemented[12] in a scenario-specific context. As input we used a collection of PDF documents provided on the website of the *Ministry for Climate Protection, Environment, Agriculture, Nature Conservation and Consumer Protection of the German State of North Rhine-Westphalia*. These kinds of documents should be relevant for the local level in Germany (Cp. deliverable D2.3, p.).

This content of the particular HTML pages was converted into RDF format, i.e. into a list of RDF resources (one resource per each PDF document) that include the title of the document, its subjects (taken from the websites), its publication year, its abstract, its source and a link to the actual PDF document. Since the purpose of the demonstrator was to present the output data of the technical conversion process and to provide a technical interface for accessing the data, this data set has been converted into RDF format manually. For the RDF representation of the metadata of the PDF documents from the HTML pages and the links to the particular PDF files, classes and properties of the Dublin Core Metadata Initiative were applied, which are commonly used for modelling documents (Cp. Chapter 3).

Technically, the output data is stored in a database and is accessible via a HTTP protocol. In this concrete example, we used the software Sesame[13], which is a framework for storing and maintaining RDF data. The data is stored in a relational database and is accessible via a SPARQL endpoint provided by Sesame, which is a specific HTTP protocol for querying RDF data. Additionally, the converted data is available as HTML representation using the tool Pubby[14].

Figure 11 shows a screenshot of the RDF prototype. This HTML representation displays the underlying RDF data in a structured way. This screenshot in particular shows an overview of all transformed PDF documents from the ministry. However, this view is not intended to be used by end-users, since the terms and links are not self-descriptive. For each transformed PDF document, there exists one resource containing the transformed metadata like title, of the document, abstract, publication date and others, and a link to the specific PDF file. This is shown in Figure 12.

[12] Cp. URL: http://lod.gesis.org/sense4us/ (Retrieved on 08/10/2014).

[13] Cp. URL: http://rdf4j.org/ (Retrieved on 11/03/2014).

[14] Cp. URL: http://wifo5-03.informatik.uni-mannheim.de/pubby/ (Retrieved on 08/10/2014).

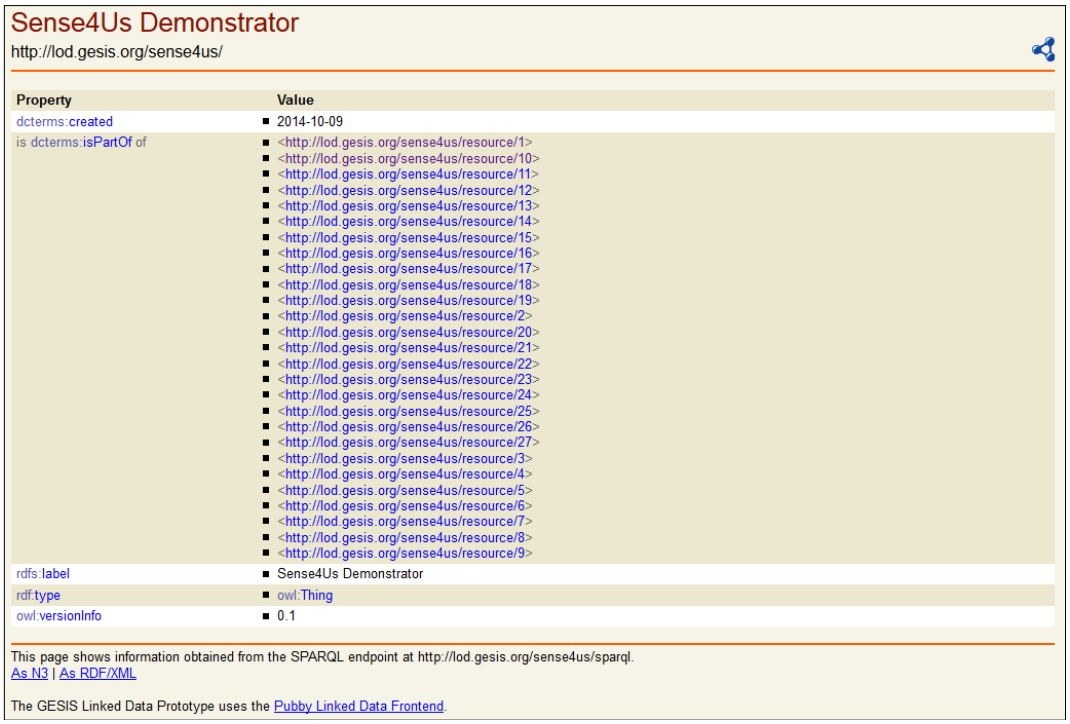

Figure 11: RDF prototype – overview of resources

Figure 12: RDF prototype - single PDF resource

Resulting from the implementation of the strategy, we can provide the following recommended actions[15] for data publishers. Knowing the underlying infrastructure on how their data is currently published, one has to determine which will be the input for the RDFization process and where the output will be stored in order to conduct the interlinking process. While for the latter aspect various state-of-the-art solutions exist like the mentioned Sesame framework, the first aspect, the input data, and needs detailed consideration. Dependent on whether the data can be retrieved via an API or web service or whether an export of the data will be used, different existing state-of-the-art methods can be used. It has also to be considered in which format the input data is available, e.g. JSON, CSV, XML. This again influences the choice of the transformation solution.

[15] These are initial investigations regarding the transformation of non-RDF data to RDF. Final recommended actions will be presented in Deliverable 4.2.2.

5 Data Interlinking on the Web of Data

Links between resources of different datasets are the key enablers of data integration on the Web of Data. These (directed) links connecting resources such as events, persons, locations or images, are RDF statements that explicitly state the relationship between the connected resources. Links may indicate equivalence (via the owl:sameAs predicate), or any other type of relationships such as domain-specific (e.g. foaf:knows, lode:atPlace). Figure 13 shows some examples of links between resources of different datasets on the Web. Since links are RDF statements, they may be stored in the source dataset (D1), together with the rest of the RDF data. Additionally, data publishers may add their links to centralized link storage and lookup services like sameAs.org[16]. This would facilitate consumer applications to find and query the links.

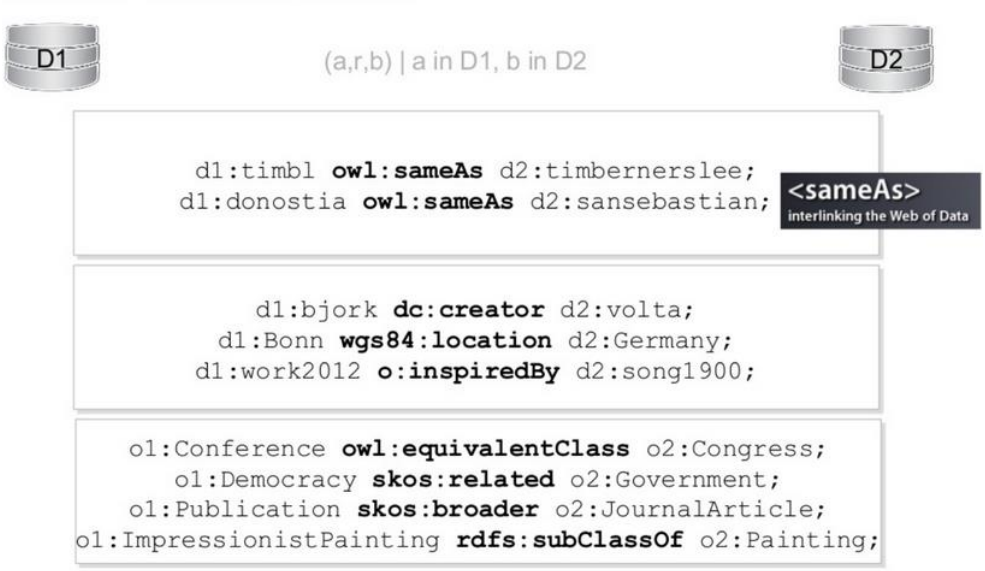

Figure 13: Examples of links on the Web of Data

Following the definitions that have been adopted in the fields of ontology alignment and instance matching [Shvaiko et al., 2013, Isele et al., 2011], the interlinking of two data sets can be defined as follows:

> *Given two data sets (D1, D2) containing the RDF description of the sets of resources A and B respectively, and R an arbitrary relation, the interlinking of A and B is the set of correspondences I_R A B such that for each correspondence the relation holds I_R = { (a,b) | aRb, a in A, b in B }.*

A data publisher can manually create the interlinking of datasets. However, since RDF datasets may contain the descriptions of thousands of resources, a manual approach does not scale. In order to be able to generate the links in a scalable manner, data publishers can use link discovery frameworks, which automatically compare the description of resources of both datasets and determine which resources should be connected, and which not.

Despite the advances in ontology alignment and link discovery technology, human intervention remains a core aspect of the process. Human input is required both for providing reference knowledge to the link discovery algorithms (i.e. configurations for semi-automatic

[16] URL: http://sameas.org/ (Retrieved on 15/03/2015).

approaches or training data for machine learning-based solutions) and for validating and enhancing the results computed by the automatic solutions. When having human resources dedicated to this task becomes unfeasible, we propose to use microtask crowdsourcing, as it provides an economic and scalable way to systematically gather the human computation required in the link discovery process. The goal of our research in the context of SENSE4Us is to investigate how to design an effective hybrid system powered by crowd workers that helps in improving purely automatic knowledge integration techniques for the Web of Data, generates valuable and good quality interlinks for policy makers and reduces the overload that data publishers may encounter in the quality assurance stage. Hence we build upon the existing work presented within the following section 5.1)

5.1 Existing approaches for knowledge integration on the Web of Data

We describe a set of relevant approaches in the context of knowledge interlinking,

Ontology alignment

After a decade of research on ontology alignment, a large variety of techniques have been defined [Shvaiko et al., 2013]. Ontology alignment and instance matching algorithms usually have three different approaches to compare entities: lexical analysis, structural analysis and semantic analysis. In lexical approaches, the values of representative labels are compared (i.e.. labels specified with properties like rdfs:label, rdfs:comment and skos:prefLabel) and the distance between labels is measured with standard measures such as Jaccard, Levenshtein or Jaro-Winkler distances. Structural analysis entails the comparison of the structure of the entities (i.e. their subclasses and superclasses, siblings and mapped entities). While semantic analysis requires the use of reasoning mechanisms and deduction of new assertions.

Prompt [Noy et al., 2003] was one of the first systems for ontology merging, alignment and versioning, based on lexical and structural analysis of ontology concepts and properties. Designed as semi-automatic tool requiring configuration by the user, it was developed for the Protègé ontology editor. LogMap [Jimenez et al., 2011] analyses the entities of pairs of ontologies on the three levels. First, it performs a lexical comparison of the labels, which may be supported by external sources like WordNet or UMLS. Second, it performs structural analysis of resources, considering descendants, ancestors and topological order. Third, anchor mappings are derived, semantic conflicts are repaired and new mappings tried to be discovered. AgreementMaker [Cruz et al., 2009] is able to generate n:m mappings by running lexical analysis using measures like TF-IDF and external sources like WordNet, as well as performing a structural similarity analysis based on descendants and siblings. HCM, developed in the context of vocabularies within the Linked Data space, is able to perform ontology alignment taking into account hundreds of ontologies at the same time [Gruetze et al., 2012].

Data interlinking

While several of the ontology alignment tools have been extended to support instance data interlinking (e.g. LogMap), there are data interlinking frameworks which have been specifically developed for the context of Linked Data. Given that purely automatic tools were not able to perform well enough and considering the difference in the size of the data (i.e. vocabularies vs datasets), several authors created batch-style semi-automatic tools that generate the interlinking between two RDF datasets following link specifications that expert users must encode manually (e.g. in XML). Silk [Volz et al., 2009] is a prominent example of such tools, which may also be configured via its graphical user interface. Once data publishers specify the type of link to be created, the type of resources to be connected, the property paths that should be compared and the string comparison measures and thresholds to be

used, Silk is able to create the RDF links that follows the link specification. LIMES [Ngonga et al., 2011] is similar to Silk, as it also requires the specification of a configuration file including the definition of the source and target data sets. LIMES uses triangular inequality techniques to reduce the number of elements to compare. MeLinDa [Scharffe et al., 2011] builds on top of Silk´s syntax, and during the analysis of instances to be linked, MeLinDa takes into account information on existing mappings between ontology entities. The major problem that data publishers face with these domain-independent and general purpose interlinking tools is that connecting resources that do not follow a perfect string match in an accurate way becomes very difficult. (Semi-)automatic tools cannot cover all irregular cases, therefore, data publishers need to go through the data to be able to publish an interlinking of higher quality.

Human computation-based interlinking

The acquisition of human knowledge relies to a large extent on the motivation mechanism implemented by the platform where the users are expected to contribute. Human Computation has tried to give incentives to non-expert users in order to create content annotations, often following the design of a Game with a Purpose [von Ahn, 2006], to motivate people by means of fun and competition. In the context of Semantic Web annotations, the INSEMTIVES project obtained remarkable results such as SeaFish, an example for collaborative semantic annotation and interlinking. SpotTheLink [Thaler et al., 2012] focuses on the task of ontology alignment. However, as argued in the tests done with SpotTheLink, games can be considered an intellectual challenge or even boring. Crowdsourcing is at the moment one of techniques that gives the most straightforward reward to requesters, and motivates workers through social contact or money among others [Kaufmann et al., 2011]. The database community has researched similar areas (entity resolution, record linkage and deduplication in databases) with microtask crowdsourcing. [Wang et al., 2012] presented a hybrid solution for entity resolution, combining human and machine computation. While the database and the Semantic Web scenarios share many commonalities, in the latter we find different additional challenges: for example, the data sources to be connected may follow different schemas and contain diverse domain information, the relations to be defined can be of any nature and the sources can be more dynamic. ZenCrowd [Demartini et al., 2012] uses microtask crowdsourcing for assigning URIs to entities that are discovered in textual Web pages, using the LOD cloud. ZenCrowd was later on extended for instance matching, but the focus of that research was not to study the interaction with automatic tools. Our initial contribution on the use of microtask crowdsourcing for ontology alignment (CrowdMAP) [Sarasua et al., 2012] lays the basis of our current work.

5.1.1 The benefit of interlinking for Sense4Us end-users (policy makers)

While (crowdsourced) interlinking technology is designed to be used by data publishers who bring their data to the Linked Data space, the resulting interlinked datasets have a positive impact on the Sense4Us use case and end-user. The main advantage of interlinking the data is that it offers an integrated space of information, which in the end allows users to execute richer queries over aggregated heterogeneous data. As an example, one could think of the Sense4Us end-user (i.e. policy maker) who is interested in better understanding why his/her country has decided to increase the investment in wind power. The PDF uploaded to the system is a document about EU budget for wind power energy from 2000 to 2020. The list of data sets identified by the relevance component are: **d1**: an excel sheet reporting the money each European country has invested in wind power from 2000 to 2014, **d2**: a .csv file with the information about companies involved in the wind power industry in each EU country, and **d3**: the LOD data set DBpedia

After RDFizing d1 and d2, the semantic publishing and interlinking components described in this deliverable could interlink d1, d2 and d3 by defining owl:sameAs links based on the (geo)location. That is, the names of (some) countries listed in the excel sheet about the investment in wind power (d1) match the names of (some) countries listed in the .csv file about the companies working in the wind power industry (d2) and countries in DBpedia.

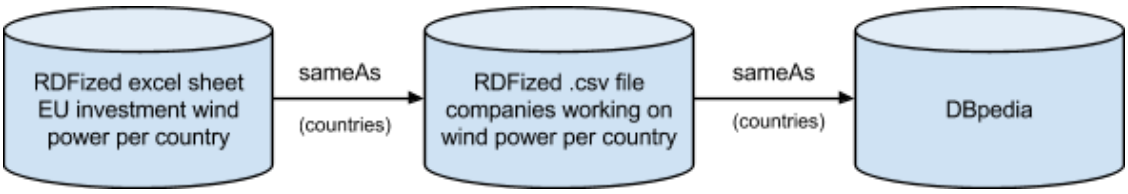

Figure 14: Interlinking example

Once the datasets are interlinked in this way, the user can now ask the system for the list of countries with a population bigger than X (from d3), and with more than 50 companies in the wind power industry (from d2), which had an increase of 50% in the investment in wind power in the last 5 years (from d1). The information comes from different data sets, and can be retrieved thanks to the relations (or links) that we define between data sets.

5.1.2 Microtask crowdsourcing

Jeff Howe defined crowdsourcing as "the task of taking a job traditionally performed by a designated agent (i.e. usually an employee) and outsourcing it to an undefined, potentially large group of people (i.e. the crowd) in the form of an open call"[17]. In Microtask crowdsourcing, workers accomplish microtasks, which are small tasks that can be done independently. When the problem to be solved is big, the requester must decompose it into small parts and publish a set of microtasks that will be run in parallel. Common microtasks include image labelling and classification, sentiment analysis of Web content, search relevance assessment, natural language translation and data validation and cleaning.

A so-called online labour marketplace is an online site where workers can browse, apply for, and work on available microtasks, which have been directly or indirectly posted by requesters. Amazon Mechanical Turk[18] is an example of such labor marketplaces, including hundreds of thousands of online microtasks. In order to ensure good quality results, MTurk calculates reputation values for workers based on their previous performance, so that requesters can use the information to reject responses from bad workers. CrowdFlower[19] is a crowdsourcing platform that acts as an intermediary between the requester and a labor marketplace. Since CrowdFlower can publish microtasks through different channels (e.g., Clickworker, Crowd Guru and Getpaid), its workforce is considered to be bigger than single marketplaces. CrowdFlower implements quality assurance methods based on gold standard evaluation, which help determine the reliability and performance of workers, and to filter spammers at run time.

[17] URL: http://archive.wired.com/wired/archive/14.06/crowds.html (Retrieved on 15/03/2015).

[18] URL: https://www.mturk.com (Retrieved on 15/03/2015).

[19] URL: https://www.crowdflower.com/ (Retrieved on 15/03/2015).

There are several issues in microtask crowdsourcing that remain research challenges. For instance, running complex work through simple microtasks is not yet a settled issue. New approaches like CrowdForge [Kittur et al., 2011], which by using coordination and distributed computing techniques is able to divide a big problem into several tasks and handle the dependencies between them. Another open issue is quality assurance. Approaches like TurKit define alternative solutions to gold standard quality control, by enabling iterative workflows, in which workers can evaluate other worker's output [Little et al., 2009]. Identifying the best match between worker and microtask can also lead to an improvement of the quality, and this is what is proposed as a task scheduler by Khazankin et al., or as a task recommendation system by Ambati et al.

5.1.3 Crowdsourced Interlinking

We define the approach of crowdsourced interlinking as the process of outsourcing the discovery of links to the crowd available at online labour marketplaces. In this scenario, the crowd does not replace (semi-)automatic techniques, but instead, it provides the support these techniques require from humans.

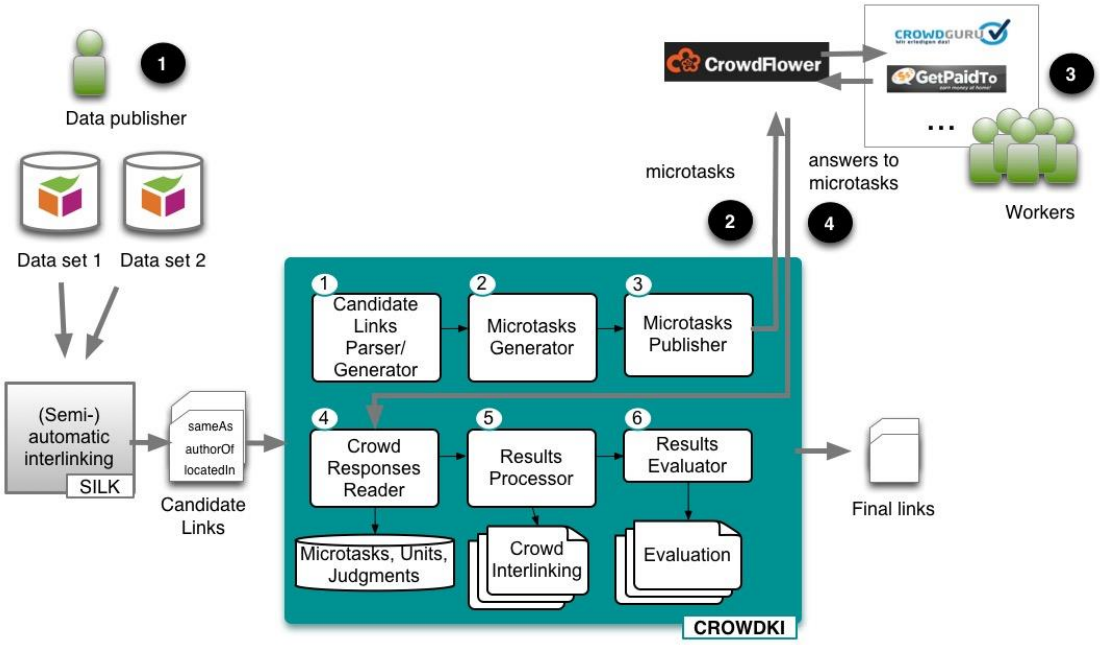

Figure 15: Overview of the process of crowdsourcing interlinking

CROWDKI (Crowd-powered Knowledge Integration) is the software that we developed as a proof of concept to investigate the approach. Given a set of links between resources of different data sets, CROWDKI takes such set of links as input, and generates microtasks that are published on marketplaces. The purpose of the generated microtasks is to make the crowd review the set of links. Crowd workers can review the links included in the published microtasks in return for a small amount of money, and once a satisfying number of trusted workers provides their responses, CROWDKI builds a set of verified links based on the aggregation of such responses.

Figure 15 depicts the process in detail. Imagine a data publisher, Sam, using Silk for the interlinking of two data sets, which have been published following the Linked Data principles. Sam configures Silk in a way that it compares resources of both data sets in order to establish sameAs and other domain-specific relationships. After its execution (see number 1), Silk

provides two sets of links: first, the set of links, which were identified with enough confidence. Second, the set of links where the tool could not establish with enough confidence whether the links should exist or not. Both sets are merged and given to CROWDKI. It is then when the links are parsed, and using the predefined UI templates and configuration options, CROWDKI creates individual microtasks which include information and questions that enable humans to review the links. CROWDKI adds some additional links from an internal set of gold standard links to enable the detection of trusted and untrusted workers. The generated microtasks are sent to CrowdFlower (see number 2), which acts as an intermediary between CROWDKI and the online labour marketplaces. Besides the fact that CrowdFlower is able to handle the gold standard links to reduce the amount of spam, the main advantage of using CrowdFlower is that it enables CROWDKI to publish microtasks in several marketplaces simultaneously (see number 3). CrowdFlower retrieves the responses of workers, and sends them to CROWDKI (see number 4). CROWDKI reads and processes the responses, keeping a local copy of the results, and subsequently generated the crowd interlinking and serializes into RDF files. Such links are considered to be the final links. CROWDKI also provides some methods to evaluate the results comparing them to a gold standard. The code, together with some technical information can be found in the CROWDKI GitHub repository[20].

5.1.4 Preliminary experiments

As a first step, and to explore the challenges we face in such scenario, we conducted a set of initial experiments in online labour marketplaces. We confirmed the feasibility of the approach, by using state-of-the-art Linked Data interlinking technology and RDF data.

5.1.4.1 Data sets and ground truth

For our experiments we selected data available at the LOD cloud and data offered by the ontology (and instance data) alignment evaluation campaigns[21]. We decided to analyse the performance of workers in common-knowledge and easy to process domains (like music, or news) first, because the chance to find crowd workers who can give reliable answers in these fields is higher than in other more specific domains (like biomedicine or the social sciences). Considering the international setting that microtask crowdsourcing provides, we only selected resources described in English.

- The **EventMedia** data set[22] contains RDF data about events and media, which was collected and triplified from other non-RDF sources like Last.fm, or Flickr. This data set is linked to DBpedia in events (owl:sameAS links), people (lode:involvedAgent links) and locations (lode:atPlace links). The existing links available in the dataset have been generated by a domain-specific interlinking tool developed by the data publisher, and the resulting links have been reviewed by the data publisher. Therefore, we considered the links in the dataset to be a ground truth we can compare with.
- The **New York Times** (NYT) data set[23], contains RDF data of subject headings generated by the American newspaper. It describes people, organizations, places and

[20] URL: https://github.com/criscod/CROWDKI (Retrieved on 15/03/2015).

[21] URL: http://oaei.ontologymatching.org/ (Retrieved on 15/03/2015).

[22] URL: EventMedia data set http://datahub.io/dataset/event-media (Retrieved on 15/03/2015).

[23] URL: New York Times Linked Open Data set http://datahub.io/dataset/nytimes-linked-open-data (Retrieved on 15/03/2015).

descriptors that were used to tagged news. The data set is linked to DBpedia in people, organizations and locations with owl:sameAs links. The interlinking was defined manually by the data publisher and it has been used in ontology alignment evaluation initiatives, thus we consider it to be a ground truth or reference interlinking to compare with. Furthermore, the OAEI evaluation campaign certified the NYT interlinks as reference interlinking between the NYT and DBpedia.

- The **persons** datasets[24] was provided by the OAEI evaluation campaign to assess instance matching algorithms. The person11 and person12 dataset (to be connected), contain personal information describing persons by the name and surname, age, date of birth, address etc. The second dataset contains the information of the first dataset with some modifications. We used the ground truth provided by the OAEI as reference interlinking to compare with.

5.1.5 Methodology

To test the accuracy that can be obtained crowdsourcing the process of instance data interlinking in the scenario that we envisioned, we defined two kind of experiments: On the one hand, we analyzed how the purely crowdsourced interlinking compares to current data interlinking technology. On the other hand, we created microtasks that presented workers with the resulting links of an interlinking tool, in order to check whether crowdsourced interlinking is able to improve its results. We gave the crowd the list of links that the tool was able to identify with high and low confident. This way, we could have the chance to improve both precision and recall-because only links identified with high confidence are considered to be accepted links.

We used Silk[25], a domain independent state-of-the-art interlinking tool. The measures that we used to evaluate the results were standard information retrieval measures: precision (to measure correctness) and recall (to measure completeness).

We designed our microtasks as simple tasks in which crowd workers need to specify whether there is a relation between the two resources being described. There are not strict design conventions around the price to pay per data unit, the number of units to include in one page and the number of workers to assign to in the microtasks. However, following evolving recommendations by the crowdsourcing research community and our own experience after working with CrowdFlower, we included few links per page (4 and 7 in the tests with the EventMedia and NYT, and 5 in the subsequent tests with the person11-person22). In the initial tests we payed $0.04 per page, and we increased this in the second tests to $0.01 per link. There are studies indicating that higher rewards do not achieve better quality, but only the speed of the results is increased. In the first round of experiments we asked for 5 different responses (of different workers) to be able to have a majority vote answer, and we decreased this number to 3 in the second round of experiments, because the quality control provided by CrowdFlower ensures good quality with 3 assignments per data unit - it is always a trade-off between the amount of money to be spent and the quality to be achieved.

CrowdFlower keeps track of the performance of crowd workers as they work on different jobs, and classifies crowd workers in three different levels of performance. When we defined the interlinking task configuration in CrowdFlower we had the opportunity to select between high quality or high speed. The first option selects crowd workers with the highest level of performance, while the second option selects crowd workers with the lowest level of

[24] URL: http://oaei.ontologymatching.org/2010/im/ (Retrieved on 15/03/2015).

[25] URL: http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/ (Retrieved on 15/03/2015).

performance. We selected the highest quality option, because our scenario aims at quality enhancement.

5.1.5.1 EventMedia and NYT toy experiments

In the first experiments, we selected for each of the EventMedia-DBpedia and NYT-DBpedia pairs of datasets a subset of their reference interlinking selecting 10 correct links and 10 incorrect links. Parsing the generated reference interlinking files, we programmatically queried the datasets and retrieved the descriptions of each of the resources involved in the reference interlinking files. This way, we created subsets of the complete datasets.

5.1.5.2 Person11-Person12 experiment

In this case we used the complete dataset. We defined the Silk configuration as follows (Figure 16): we indicated the two data sources to be connected, restricted Silk to connect only resources typed as Persons, and defined an aggregation of comparisons for the linkage rule (e.g. combining the JaroWinkler distance between the names of persons, and the numerical comparison of the age of the persons). We defined a threshold of 0.85 (i.e. resources with an average comparison value under 0.85 should not be linked) and set the confidence limit to 0.95 (i.e. pairs of resources for which Silk does not compute a confidence higher than 0.95 are not declared as links). We retrieved the RDF statements with the accepted links in a file, and separately we collected the pairs of resources for which Silk was not confident enough but showed a considerably high comparison value). We crowdsourced both sets of links (accepted links and to-be-verified links) together. This way had the chance to improve precision and recall of the links that Silk declared as links (i.e. only the accepted set).

```
<DataSources>
  <DataSource id="s1" type="file">
  <Param name="file" value="person11.rdf" />
  <Param name="format" value="RDF/XML" />
</DataSource>
<DataSource id="s2" type="file">
  <Param name="file" value="person12.rdf" />
  <Param name="format" value="RDF/XML" />
  </DataSource>
</DataSources>

<Interlinks>
  <Interlink id="persons">
    <LinkType>owl:sameAs</LinkType>

    <SourceDataset dataSource="s1" var="a">
      <RestrictTo>
        ?a rdf:type p1:Person
      </RestrictTo>
    </SourceDataset>

    <TargetDataset dataSource="s2" var="b">
      <RestrictTo>
        ?b rdf:type p2:Person
      </RestrictTo>
    </TargetDataset>

     <LinkageRule>
     <Aggregate type="average">
<Compare metric="jaroWinkler" required="true">
     <Input path="?a/p1:given_name" />
     <Input path="?b/p2:given_name" />
   </Compare>
   <Compare metric="jaroWinkler" required="true">
     <Input path="?a/p1:surname" />
     <Input path="?b/p2:surname" />
   </Compare>

  <Filter threshold="0.85" />

  <Outputs>
    <Output type="file" minConfidence="0.95">
      <Param name="file" value="accepted_links.nt" />
      <Param name="format" value="ntriples" />
    </Output>
    <Output type="file" maxConfidence="0.95">
      <Param name="file" value="verify_links.nt" />
      <Param name="format" value="ntriples" />
    </Output>
  </Outputs>
```

Figure 16: Example RDF/XML notation

This link specification leads to a set of 161 links accepted by Silk and 279 links to be verified (and therefore not being declared as links). In order to have quality assurance in CrowdFlower we generated additional links (built from real resources of the datasets) and programmatically introduced their ground truth label (i.e. correct answer) in the CrowdFlower. This way, by combining the test links with the "normal" links that we need the crowd to process, CrowdFlower checked the accuracy of workers. We defined as a requirement that our trusted workers must have at least a 70% accuracy in test links.

The following two figures show the user interface of two of our microtasks:

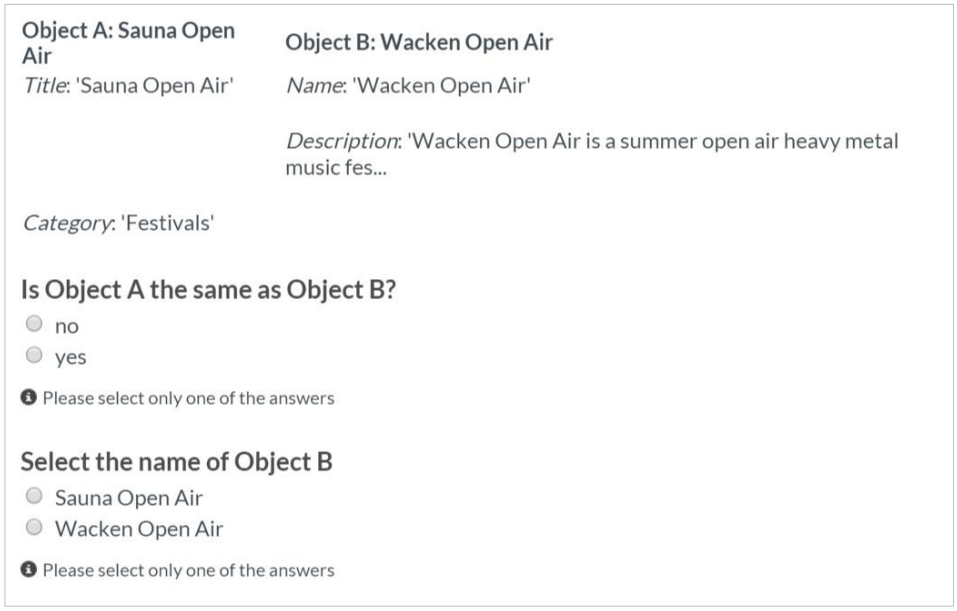

Figure 17: User interface 1 – Interlinking microtasks

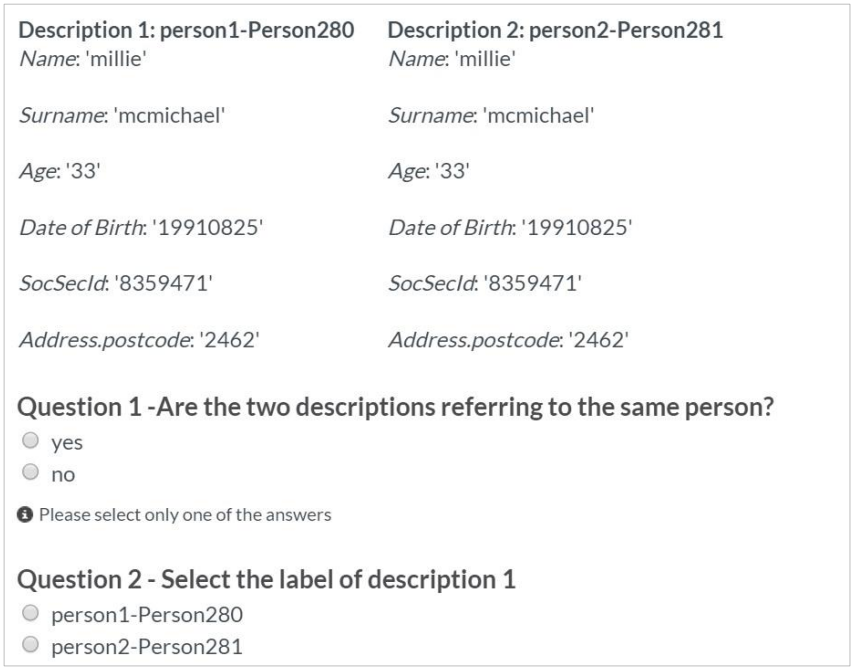

Figure 18: User interface 2 - Interlinking microtasks

5.1.6 Results

The following two tables show the results in the experiments where we studied the feasibility of the approach with the EventMedia and NYT datasets. The values for the precision (p) and recall(r) show that CROWDKI is able to review links with the crowd successfully.

	EventMedia - DBpedia events	EventMedia - DBpedia people	EventMedia - DBpedia locations	NYT - DBpedia people	NYT - DBpedia organizations	NYT - DBpedia locations
Silk t=0.85, c=0.95	p=1.0 r=0.3	p=1.0 r=0.9	p=1.0 r=0.6	p=0.05 r=0.0	p=1.0 r=0.6	p=1.0 r=0.3
CROWDKI	p=0.91 r=1.0	p=0.9 r=0.8	p=1.0 r=0.6	p=0.91 r=1.0	p=1.0 r=1.0	p=0.59 r=1.0

Table 4: Accuracy of the crowd in crowdsourced interlinking – Input half correct links, half incorrect links

	EventMedia - DBpedia events	EventMedia - DBpedia people	EventMedia - DBpedia locations	NYT - DBpedia people	NYT - DBpedia organizations	NYT - DBpedia locations
Silk t=0.85, c=0.95	p=1.0 r=0.3	p=1.0 r=0.9	p=1.0 r=0.6	p=0.05 r=0.0	p=1.0 r=0.6	p=1.0 r=0.3
Silk+CROWDKI	p=1.0 r=0.3	p=1.0 r=0.8	p=1.0 r=0.3	p=0.91 r=1.0	p=1.0 r=0.91	p=1.0 r=0.3

Table 5: Accuracy of the crowd in crowd sourced interlinking – Input result of Silk

What can be interpreted from the results is that CROWDKI can be useful to improve the recall of Silk, since Silk already performs with a good precision - a result that may vary depending on the data. At the moment, with the workflow we defined, the only chance to improve the recall from the result of Silk is asking the crowd about the links that Silk generated with high confidence - included in the verified links file. This suggests that there is room for researching new techniques to retrieve potential candidates for the interlinking that could help in improving the recall of Silk.

CROWDKI showed low accuracy in the two cases where the workers were presented with location information. We showed the coordinates with numbers, and this might have confused some of the workers since the numbers for latitude and longitude were on purpose quite similar. In future experiments we will include location information within a map. Some of the cases that made the precision of Silk decrease were instances that contain very similar names, like for example ``John F. Kennedy Jr" and his father. Also, ``Apache Corp." and ``Apache Software Foundation". This shows that humans sometimes also do not achieve perfect precision and recall.

We also experienced the quality issues present in LOD data, which make the data hard to be reused. For example, there are values that contain HTML tags within the text, and this is something that needs to be cleaned before reuse.

The results of the people's dataset are presented in the following table:

Person11-Person12	*Precision*	*Recall*
Silk	1.0	0.322
Silk + CROWDKI	1.0	0.878

Table 6: Experiment results

Since we set the comparison threshold high (0.85 averages of all attribute comparisons) and we required Silk to have a high confidence (0.95) to declare pairs of resources to be links, we had as a result high precision. A positive result is that the crowd was able to reach the same perfect precision. On the other hand, Silk had a low recall, which was considerably improved by the crowd. It was valuable to have the crowd review all the links that Silk was not able to decide about confidently.

In some of the pre-tests that we did for this data, we observed how difficult it is to decide on labels without contextualizing the problem at hand correctly. In one particular microtask the crowd was asked whether two descriptions of persons with the same name (and other equal attributes) but with different date of birth, were referring to the same person. Some crowd workers were saying that they were not about the same person, because the date of birth was not the same. However, in the Web of Data scenario, we can encounter such case if one of the descriptions were generated with data quality issues (e.g. typos). This was exactly the case, because one description showed a date year0203 while the other description showed year0302. If we consider the descriptions to be probably incorrect we can decide that they talk about the same person, but if we trust the descriptions completely and assume they are correct, we would say they are not about the same person. What we did was to reformulate the instructions of our microtasks to explain this better.

The microtasks were completed in a couple of hours and we collected 1.343 trusted judgments (i.e. labels for links). According to CrowdFlower's computation of the quality of the job (by selecting 100 random links processed by the crowd) we had a 99.24% agreement - measured by counting at the number of crowd workers who agreed with the majority vote.

We had 109 contributors reviewing links coming from 9 different marketplaces. But only few contributors processed a large amount of links. Even if the total number of links in the job was 490 links (440 normal links + 50 test/golden links), we had restricted the maximum number of links per person to 250 (the default value suggested by CrowdFlower). The following figure shows the number of judgments (i.e. links processed) per contributor, for the top 100 crowd workers who participated in our tasks. The behaviour is normal in the context of microtask crowdsourcing: tasks are simple but also repetitive, so crowd workers get bored at some point and switch the tasks to work on. However, this is convenient for our knowledge acquisition process, because this way we can have the opinion of more people. It would not be reliable to have only 3 persons labelling the whole dataset.

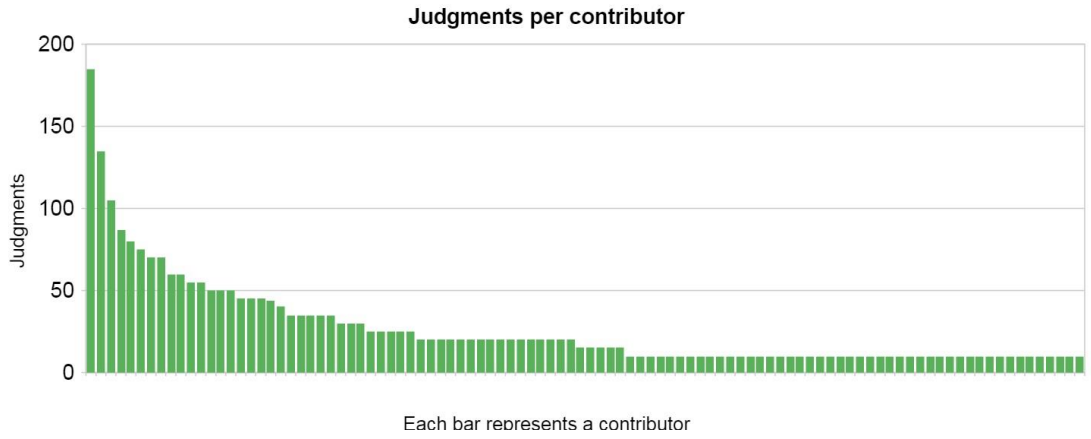

Figure 19: Judgements per contributor

Table 7 shows several lessons from experiments in online labour marketplaces:

Lessens from experiments in online labour marketplaces (Cp. section 5.1.4)

• First, there is not always the need to include a crowd-powered layer to the interlinking process. If we think of the SENSE4US use case, when we connect datasets that have been geo-located and include a standardized ISO code for countries Silk would perform perfectly because the codes can be perfectly matched by string comparison. It would make no sense to spend human and economic resources in reviewing the data. The crowd-powered solution becomes really valuable in cases in which the data is very heterogeneous, or it contains quality problems.
• Second, when we crowd source data processing to the wide audience, it is important to:
  ○ give clear instructions about what we expect the people to do;
  ○ present the data in an easy to understand and attractive way; and
  ○ avoid technical terms like “Ontology” or “RDF”, because they only add confusion to people with no literacy in Semantic Web.
• Third, the potential enhancement of the interlinks will depend on the amount of contextual information present in the description of the resources. If persons were only labelled with a name, crowd-workers would also have difficulties in deciding whether the two persons are the same or not, due to a lack of information.

Table 7: Lessens from experiments in online labour marketplaces

5.1.7 On-going and future work

On-going and future work focuses on the optimization of the approach described above. First, we will study the approach in a different scenario: an Active Learning-based interlinking process, in which the configuration of link specifications is no longer required to be generated by an expert user (i.e. the data publisher). Crowd workers would label sampled data, and the algorithm iteratively and interactively learns the link specifications, which afterwards are executed by the interlinking engine (e.g. Silk) to generate the resulting links. State-of-the-art interlinking frameworks like Silk and LIMES have been recently extended to incorporate Active Learning algorithms. Second, we will analyze whether a task assignment (or crowd-link selection) method would have any influence in the performance of approach. We will continue experimenting with more data. One possibility could be to use vocabularies and data of digital libraries, as we discussed in the context of the talk we gave at the Semantic Web in Libraries conference (SWIB2014)[26].

[26] URL: http://de.slideshare.net/cristinasarasua/swib2014csarasua (Retrieved on 15/03/2015).

6 Crowd Work CV

One of the challenges in human computation systems is to involve the humans who, as intelligent and independent beings with a particular knowledge, are crucial to solve problems that machines can hardly solve alone. Crowdsourcing alleviates this challenge, as it provides a mechanism to distribute a task among a potentially large group of people. A promising strategy to improve the quality of crowd work, which is particularly relevant for knowledge-intensive crowdsourced tasks, is to find the most suitable worker(s) for a microtask (or vice versa), as Kittur et al. (2013) highlighted. This is particularly relevant to knowledge-intensive microtasks like our interlinking tasks.

However, the realisation of such process is hindered by the current microtask crowdsourcing infrastructure, which is highly focused on independent marketplaces. Even if many of them have adopted some common patterns, each of them acts as a data silo. When crowd workers are registered and work in several marketplaces, the work they perform is registered in the marketplace they worked at and only visible there. If a requester is interested in knowing further information on the achievements and proven skills of a worker (e.g. through obtained qualifications) in other marketplaces, this information is not accessible programmatically, even though the data exists and it is visible to the worker. The same applies to requester information. This lack of data interoperability between marketplaces has a negative impact in the process of finding the best combination of workers and microtasks and may result in uninformed decisions. To overcome this limitation, we propose Crowd Work CV [Sarasua et al., 2014], an RDF-based data model to represent someone's crowd work life, equivalently to what traditional Curriculum Vitae reflect.

6.1 Existing approaches for crowd profiling and task assignment

Several authors have proposed new methods for matching crowd workers and tasks in crowdsourcing environments. Khazankin et al. (2011) defined a framework for selecting suitable crowd workers to solve a task based on skill requirements attached to tasks, the availability workers report they have, and the skills workers have. Goel et al. (2013]) introduced a method for assigning tasks to workers, which analyses both skills and costs. Difallah et al. (2013) implemented in a Facebook App a recommendation strategy that pushes suitable tasks to users based on information extracted from their Facebook profiles and previously accomplished HITs, following various assignment strategies (i.e. category-based, text-based, and graph-based). These approaches do not offer a shareable and reusable description of worker expertise that could be used across-platforms. Ul Hassan et al. (2013) proposed the SLUA ontology for matching users and actions in crowdsourcing scenarios. While the authors raised the problem of lacking interoperability between platforms aligned to our initial proposition, their approach has a different focus: they describe tasks, users, rewards and capabilities primarily for routing. In contrast, our goal is to gather more information and be able to share it as a means to recognition for work. We in addition consider microtasks, marketplaces, qualifications and requesters' information. Moreover, our data leads to a workflow for building CV summaries out of large sets of RDF triples.

ResumeRDF[27] is an RDFS vocabulary to express information of Curriculum Vitae, including personal details, attended courses, skills and work experience. Celino proposed the Human Computation ontology[28], which enables the annotation of crowd sourced data and is mapped

[27] URL: http://rdfs.org/resume-rdf/ (Received on 15/03/2015).

[28] URL: http://swa.cefriel.it/ontologies/hc.html (Retrieved on 15/03/2015).

to the Provenance Ontology. These data models share some common concepts with ours but do not cover all the crowdsourcing-specific domain required in a Crowd Work CV.

6.1.1 Recognition for Micro Work

As a motivational scenario, let us imagine Alice, who has registered in several marketplaces (e.g. ClixSense, GetPaid, Neobux, and GlobalActionCash). She is being assessed as a candidate crowd worker for a group of microtasks published at ClixSense about sentiment analysis of Spanish Web sites. The requester who published the microtasks trusts experienced crowd workers more than inexperienced crowd workers.

- Alice registered at ClixSense but did not work there yet.
- Alice worked on text translation at Neobux, where she obtained a Spanish qualification that a requester defined.
- At GetPaid Alice successfully completed several microtasks which are equivalent to those for which she is going to be assessed, because CrowdFlower distributed the group of microtasks over several marketplaces (ClixSense and GetPaid).
- At GlobalActionCash Alice worked with very good performance on other microtasks, which required her to analyse the sentiment of Tweets--similar purpose, with different type of data.

At ClixSense, Alice will be poorly evaluated because her ClixSense work history is empty. Other candidate crowd workers who have worked on Web site sentiment analysis microtasks with a much lower performance than what Alice did at GetPaid will be better considered. Alice has a language qualification and she has proven to be capable of solving the type of job being analysed, and even other related microtasks dealing with a similar problem. Still, due to a lack of shared information, the requester will not consider her work experience. This has **drawbacks for both parties**: the requester is not taking advantage of a potentially good worker for the task at hand, and the crowd worker is missing an opportunity to work on something she might be interested in because of its similarity to previously completed crowd work.

6.2 Crowd Work CV ontology

The Crowd Work CV data model describes crowdsourcing agents (i.e. crowd workers and requesters), their interests, obtained qualifications and work history. The data model, implemented as an OWL[29] ontology[30] is available online[31]. Figure 20 shows the graphical representation of the ontology.

[29] URL: OWL http://en.wikipedia.org/wiki/Web_Ontology_Language (Retrieved on 15/03/2015).

[30] URL: Ontology http://en.wikipedia.org/wiki/Ontology_(information_science) (Retrieved on 15/03/2015).

[31] URL: Crowd Work CV Ontology https://github.com/criscod/CrowdWorkCV/tree/master/ontology (Retrieved on 15/03/2015).

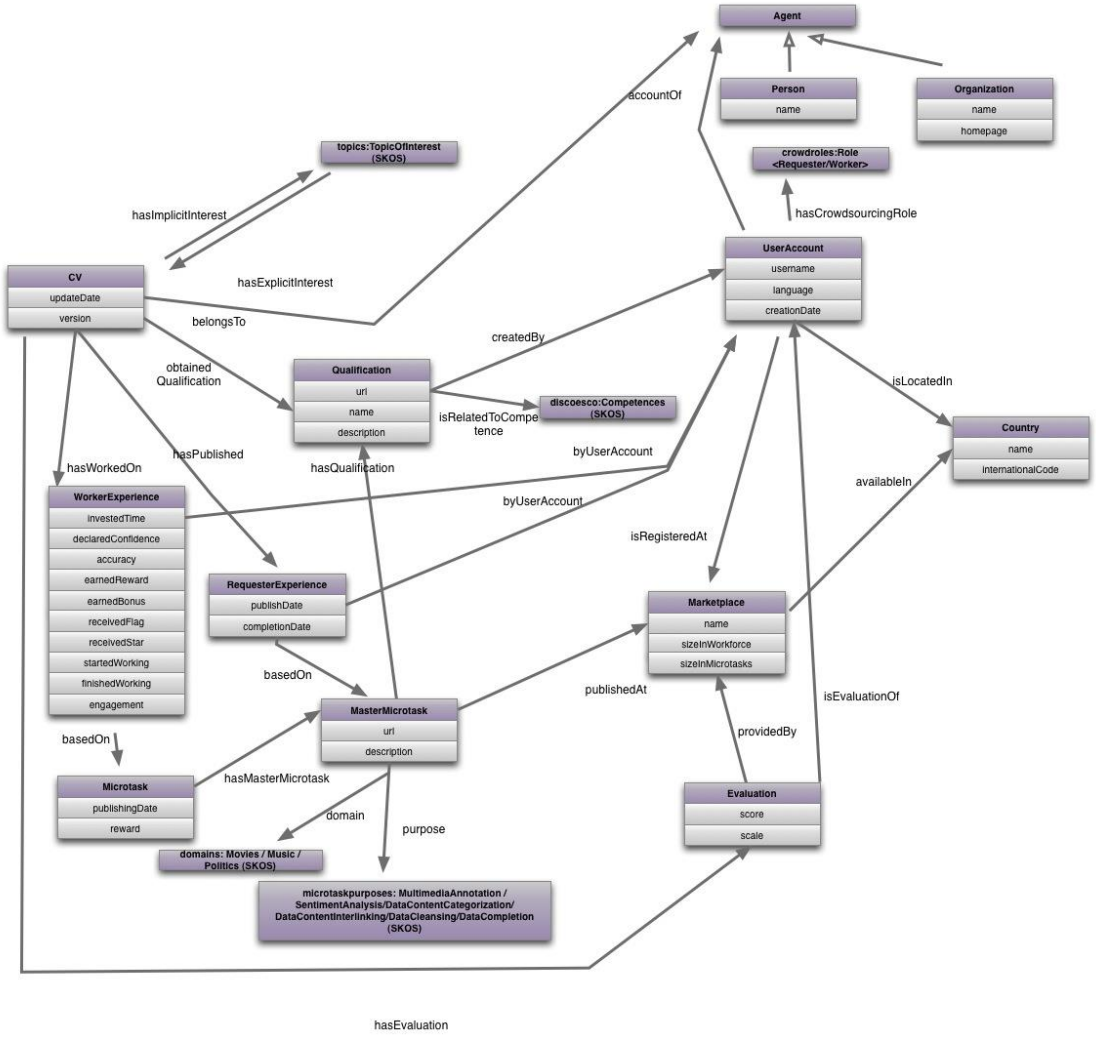

Figure 20: Overview of the Crowd Work CV ontology, for describing agents, their user accounts, CVs, qualifications, work experiences, microtasks, their master microtasks, and marketplaces.

We provide a summary of the most relevant elements of the ontology:

- CV is the core class of the ontology. It aggregates all the information that is used to report the crowd work life of an Agent, which can be either a Person or an Organisation. A CV may refer to the interests of its owner, which might have been explicitly stated by the owner (i.e. explicit interest), or might have been inferred by the interaction within the marketplace or the crowdsourcing platform (i.e. implicit interest).
- When we think of crowd workers, a CV may be related to obtained qualifications, which are related to competences (e.g. SKOS vocabulary for Europass). For each piece of work accomplished, the CV keeps an entry stated as an instance of WorkerExperience, which consists of information about the way the crowd worker solved the microtasks (e.g. the time the worker invested, whether the requester gave flags or stars in such work, and the engagement of the worker in the complete group of microtasks). The case of requesters is analogous.
- Qualification refers to the achievement that determines whether an agent (usually a crowd worker) has the required knowledge on a particular topic. Requesters can write their own tests or reuse the questions provided by marketplaces. In the Crowd

Work CV, qualifications may be defined with a textual description, a URL with a deployed example and a name. This class may be extended in the future if categories of qualifications are defined (e.g. language qualifications could be a subclass of Qualification).

- Microtask represents the particular instances of MasterMicrotasks. The specific unit of work that crowd workers need to solve. WorkerExperience is related to Microtask, since the information associated to the WorkerExperience (e.g. accuracy, invested time) is based on the results obtained in the microtasks. A Microtask is related to the MasterMicrotask which from it originated (after combining a template with data).
- Marketplace represents the crowdsourcing platform containing a Website where microtasks are offered and accomplished. Marketplaces may be described by a name and their sizeInMicrotasks and sizeInWorkforce to have some statistical information about them.
- Evaluation reflects the assessment of an Agent in a Marketplace. It is generally described by a score within a scale, but it could easily be extended, for example with new intermediate properties which, following the criteria suggested by Turkopticon, express that a Requester is evaluated by its communicativity, generosity, fairness and promptness[32].

The goal of the approach is to aggregate relevant data to describe a crowdsourcing agent's identity and enable recognition for micro work. Contributors can have a way to show what they did before and this might motivate them more. It also boosts transparency in the process of interacting with requesters, and could lead to better relations of trust. Requesters, on the one hand, can have their own CV and show workers what they offered before, to attract a regular set of contributors. On the other hand, they can use Crowd Work CV information to improve the selection processes or weight the answers they get from workers. The Crowd Work CV could also be integrated with LinkedIn information.

Crowd Work CV enables the execution of queries in a cross-platform fashion. For example, now we can answer questions like ''how good is this crowd worker in average in this type of task?'', or ''does this crowd worker have experience in this domain (e.g. music)?''. Regarding requesters, we can answer ''in which marketplace can I mainly find tasks of this requester?'' and ''for which other requesters have workers, who also worked for me, worked?''.

6.2.1 Why an RDF-based data model?

By representing the Crowd Work CV in RDF:

- we guarantee semantic and syntactic interoperability and we can ensure the specification of an agreed upon vocabulary
- we enable the integration of Crowd Work CV data with other applications, Web data sets (e.g. Linked Data cloud) and marketplaces
- we build a graph of requesters, microtasks and contributors which leads to many possibilities in terms of data and network analysis
- we define an extensible data model that enables knowledge inference

[32] Cp. Turkopticon's evaluation criteria. URL: http://turkopticon.ucsd.edu/help (Retrieved on 15/03/2015).

6.2.2 Crowd Work CV data management

We define the Crowd Work CV data management workflow (see Figure 21) as follows:

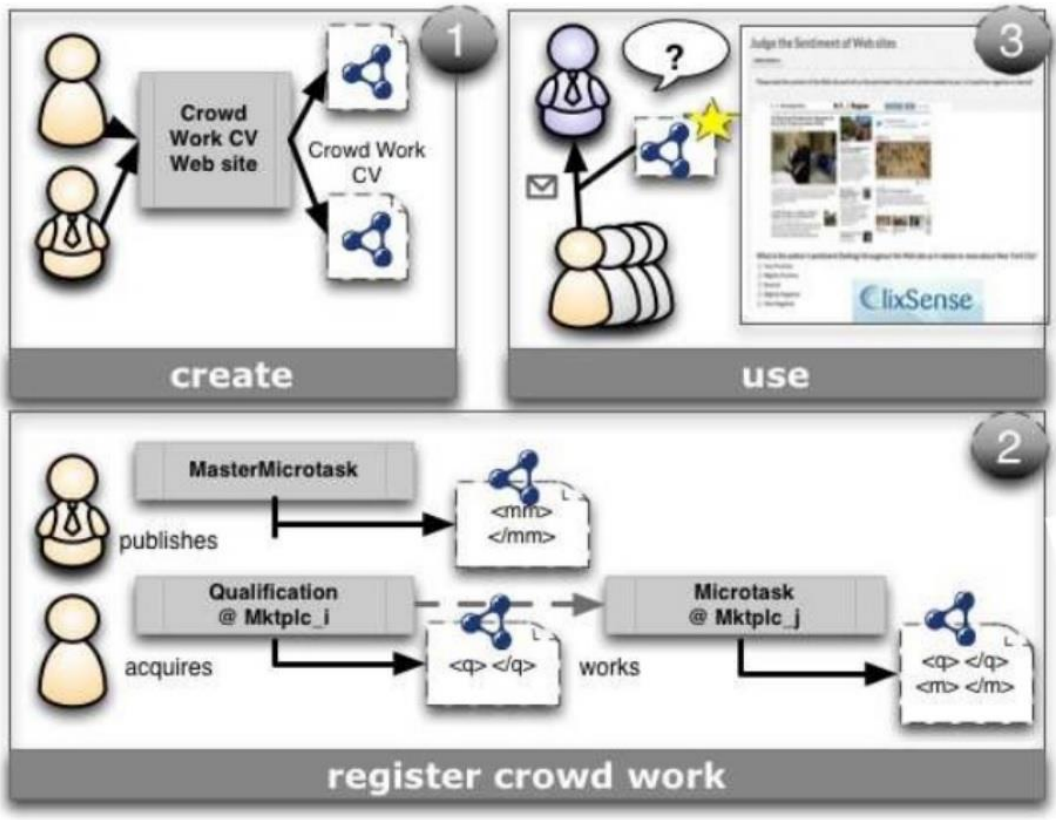

Figure 21: Crowd Work CV data management workflow

Step 1: Create

The first step of the process is to create Crowd Work CVs. Crowd workers and requesters interested in having a Crowd Work CV voluntarily register through a Web site included in the Crowd Work CV management system. Filling in a form, the system creates a new CV, as well as basic information about the owner of the CV. Alternatively, if CrowdFlower adopts the Crowd Work CV, the CV may also be created via the CrowdFlower contributor interface.

Step 2: Register crowd work

The second step is to register the work that takes place in marketplaces. When the requester publishes the microtasks, his Crowd Work CV registers the information of the published work. When Alice (the worker of the example above) works in the marketplaces the information is also updated in her Crowd Work CV. She gets the information of the acquired qualification, as well as the microtasks at GetPaid and GlobalActionCash as entries (in the form of RDF instances) in her Crowd Work CV.

Step 3: Use

Following the scenario, once the data is in the CV, Alice can use it when applying for the work offered by the requester at ClixSense. There are two substeps to take into account: first, because Crowd Work CVs can end up accumulating many entries, a representative summary of the Crowd Work CV data must be generated. A Crowd Work CV summary emphasizes relevant work patterns and it can be personalised for the ongoing job application. Second, the

Crowd Work CV system must ask the marketplaces to certify the data (to prevent having manipulated CV data). OAuth and digital signatures can be used for this purpose. Next, the requester receives Alice's application with her Crowd Work CV.

The Crowd Work CV approach does not restrict the way the data is used: it can be used in restrictive assignment strategies (e.g. someone with very low or bad experience is not allowed to work on a particular microtask), or in mixed and more open strategies (e.g. anyone can work on the microtask, but the experienced workers get a bonus or are trusted more during the aggregation of responses).

6.2.3 Crowd Work CV ontology verification

In order to ensure that we are following best practices in ontology engineering, we validated our ontology with the OOPS! pitfall scanner[33], which considers a list of 40 common pitfalls in ontology specifications. Except for the imported concepts and properties from other ontologies, we ensured that we do not have important nor critical piftalls.

We also compared the fulfillment of the aforementioned Crowd Work CV requirements: the main elements of the Crowd Work CV ontology refer to domain-independent objects in crowdsourcing systems (e.g. microtasks, user accounts and marketplaces). The ontology elements are abstract and can be used in different marketplaces. For example, the overall evaluation of a worker in a marketplace or the qualifications, do not refer to particular evaluation schemes that for example MTurk may have. The Crowd Work CV ontology can easily be extended by defining subclasses (e.g. subclasses of qualifications), subproperties, or adding new relations between existing and new ontology concepts. The SKOS vocabularies can also be easily extended in order to have for example, a broader catalogue of microtask purposes. The ontology is aligned with standard traditional CV information, describing the particular instances of work experience, the educational achievements (in our case qualifications) and related skills.

6.2.4 Future work

Future work in this direction will focus on the development of the Crowd Work CV data management infrastructure and the analysis of the impact of using the Crowd Work CV approach for microtask assignment.

[33] URL: http://oeg-lia3.dia.fi.upm.es/oops/index-content.jsp (Retrieved on 15/03/2015).

7 Publishing the interlinked Data

In general, the publication of the interlinked data means that the RDFized Open Data generated in Section 4 is published together with the links detected during the interlinking process described in Section 5. The publication of this interlinked data on a web server is the next step in the process of semantic linking and consolidation of heterogeneous Open Data. Technically, it is necessary in order to enable dereferenceable URIs (i.e. URIs that provides RDF descriptions in response to HTTP GET requests). The same technical infrastructure as described in Section 4 can be applied for this purpose.

This publication step allows for the technical connection of the RDFized data to the Linked Open Data cloud via the detected links. When a data publisher intends to be included visually in the Linked Open Data cloud diagram, the requirements[34] for the inclusion of data sets have to be fulfilled. The process of semantic linking and consolidation of heterogeneous Open Data described in this Deliverable conducts most of the required steps and leaves editorial tasks for the data publisher that have to be finished like registering and describing the data set on the portal DataHub[35].

[34] Cf. URL: http://lod-cloud.net/#how-to-join (Retrieved on 11/03/2015)

[35] Cf. URL: http://datahub.io/ (Retrieved on 11/03/2015)

8 Linked Data validation and documentation

Once the RDF data is generated, interlinked and pre-published it is important to confirm that it follows the complete set of Linked Data principles. In order to do so and following the best practices defined in the Linked Data community [Heath et al., 2011] we will:

- check that the RDF data has been serialized in a valid RDF format. We can use the W3C validator for such purpose[36].
- check that it is being offered as Linked Data. We can use the Vapour validation tool[37], which queries the data of our dataset remotely.
- we will browse the resources with a browser or visualization tool.
- we will define a set of test SPARQL queries as requirement and check with a SPARQL query engine that our dataset is able to provide correct and complete results for them.

Furthermore, in order to enable the reuse of our data by third-parties and to be aligned with best practices of Linked Data publishers, we will document our datasets properly:

- we will register our dataset in the CKAN DataHub[38] catalogue, introducing DCAT-based information.
- we will generate a voiD[39] description of the dataset indicating among other things the number of triples it contains, the number of links and representative sample resources.
- we will include the interlinks in the sameAs.org[40] service.
- we will create a user-friendly Web page to explain the acquisition of the data and any other relevant details of the publishing process.

[36] URL: http://www.w3.org/RDF/Validator/ (Retrieved on 15/03/2015).

[37] URL: http://validator.linkeddata.org/vapour (Retrieved on 15/03/2015).

[38] URL: http://datahub.io/ (Retrieved on 15/03/2015).

[39] URL: http://www.w3.org/TR/void/ (Retrieved on 15/03/2015).

[40] URL: http://sameas.org/ (Retrieved on 15/03/2015).

9 Conclusions

Within this deliverable we presented a conceptual overview of the ongoing work of WP4, how the different processes fit together and what we are planning to do in the near future. We identified different actors, who make sense of the WP4 processes or influence their outcome: the data publisher to interlink the own data with the LOD cloud to increase its visibility; the crowd to support the data publisher to interlink the data via crowd sourcing; the policy maker to run richer searches on Linked Open Data; and the project partner themselves to provide input for the LOD ranking component.

During the first year of the project we collected a couple of data sources that were identified during the end user engagement or were suggested by project partners. We then categorized the data sources to identify different aspects that are important to take into account.

Our investigation regarding the data selection within the Sense4Us scenario has influenced our conclusion on the transformation of non-RDF data to RDF data. Since Open Data is provided in a variety of different portals, interfaces and formats on the Web, we can only advice to the publishers of particular data sets how they can transform and publish their originally non-RDF Open Data in RDF format. When facing vocabulary design of data that is to be transformed, best practices of the Linked Data community should be followed like the reuse of existing vocabularies as much as possible. Additionally, an extensive data publication in RDF allows for detecting more suitable entities for the interlinking process.

Research has shown that the design of hybrid solutions, combining human and machine computation, leads to better results in knowledge curation and integration processes. That is the reason why there is a growing interest in including crowdsourcing techniques in Semantic Web tasks [Bernstein et al., 2014]. We have used microtask crowdsourcing in the context of data interlinking, one of the most critical steps towards the realization of a global data space on the Web. The preliminary results of our investigation suggest that using microtask crowdsourcing to systematically involve humans in interlinking tasks (e.g. validation of candidate links) is feasible, cost- and time-effective. One of the challenges that we will need to face in the future is the definition of new methods to optimize the process in terms of targeted contributors and tasks, without neglecting the importance of providing fair and appealing tasks to the crowd. A first step in this direction is our contribution for crowdsourcing marketplaces: the proposal of using machine-readable Curriculum Vitae to encourage recognition for micro work.

For the technical publication of the transformed and interlinked data, existing state-of-the-art solutions can be used. Which of these solutions proves to be the most suitable and feasible depends again on the data publishers and their existing technical infrastructures.

Until month 30 we will improve our initial investigation on task *T4.1 – Analysis of Open Data Sources for Relevance & Structure* and task *T4.2 – Data Mapping, Consolidation & Linking* to provide the final deliverable *D4.2.2 - Updated Investigation into Tools & Techniques for Semantic Linking of & Consolidation of Heterogeneous Open Data* regarding our work presented within this deliverable.

10 References

[1] Antoniou, G., & Van Harmelen, F. (2004). A semantic web primer. MIT press.

[2] Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, 1(1), 1-136.

[3] Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In The Semantic Web–ISWC 2014 (pp. 245-260). Springer International Publishing.

[4] Shvaiko, P., & Euzenat, J. (2013). Ontology matching: state of the art and future challenges. Knowledge and Data Engineering, IEEE Transactions on, 25(1), 158-176.

[5] Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk-A Link Discovery Framework for the Web of Data. LDOW, 538.

[6] Noy, N. F., & Musen, M. A. (2003). The PROMPT suite: interactive tools for ontology merging and mapping. International Journal of Human-Computer Studies, 59(6), 983-1024.

[7] Jiménez-Ruiz, E., & Grau, B. C. (2011). Logmap: Logic-based and scalable ontology matching. In The Semantic Web–ISWC 2011 (pp. 273-288). Springer Berlin Heidelberg.

[8] Gruetze, T., Böhm, C., & Naumann, F. (2012). Holistic and Scalable Ontology Alignment for Linked Open Data. In LDOW.

[9] Cruz, I. F., Antonelli, F. P., & Stroe, C. (2009). AgreementMaker: efficient matching for large real-world schemas and ontologies. Proceedings of the VLDB Endowment, 2(2), 1586-1589.

[10] Ngonga, A. C. N., & Auer, S. (2011). Limes-a time-efficient approach for large-scale link discovery on the web of data. integration, 15, 3.

[11] Scharffe, F., & Euzenat, J. (2011). MeLinDa: an interlinking framework for the web of data. arXiv preprint arXiv:1107.4502.

[12] Von Ahn, L. (2006). Games with a purpose. Computer, 39(6), 92-94.

[13] Thaler, S., Simperl, E., Siorpaes, K., & Wölger, S. (2012). SpotTheLink: A Game-Based Approach to the. Collaboration and the Semantic Web: Social Networks, Knowledge Networks, and Knowledge Resources: Social Networks, Knowledge Networks, and Knowledge Resources, 40.

[14] Kaufmann, N., Schulze, T., & Veit, D. (2011, August). More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. InAMCIS (Vol. 11, pp. 1-11).

[15] J Wang, T Kraska, MJ Franklin, J Feng (2012). Crowder: Crowdsourcing entity resolution. Proceedings of the VLDB Endowment.

[16] Demartini, G., Difallah, D. E., & Cudré-Mauroux, P. (2012, April). ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In Proceedings of the 21st international conference on World Wide Web (pp. 469-478). ACM.

[17] C Sarasua, E Simperl, NF Noy. (2012). CrowdMAP: Crowdsourcing ontology alignment with microtasks. Proceedings of the International Semantic Web Conference, ISWC 2012.

[18] Kittur, A., Smus, B., Khamkar, S., & Kraut, R. E. (2011, October). Crowdforge: Crowdsourcing complex work. In Proceedings of the 24th annual ACM symposium on User interface software and technology (pp. 43-52). ACM.

[19] Little, G., Chilton, L. B., Goldman, M., & Miller, R. C. (2009, June). Turkit: tools for iterative tasks on mechanical turk. In Proceedings of the ACM SIGKDD workshop on human computation (pp. 29-30). ACM.

[20] Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., ... & Horton, J. (2013, February). The future of crowd work. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 1301-1318). ACM.

[21] Cristina Sarasua, Matthias Thimm. Crowd Work CV: Recognition for Micro Work. In Proceedings of the 3rd International Workshop on Social Media for Crowdsourcing and Human Computation (SoHuman'14). Barcelona, Spain, November 20

[22] Khazankin, R., Psaier, H., Schall, D., & Dustdar, S. (2011). Qos-based task scheduling in crowdsourcing environments. In Service-Oriented Computing (pp. 297-311). Springer Berlin Heidelberg.

[23] Gagan Goel, A.K., Singla, A. (2013) : Matching workers expertise with tasks: Incentives in heterogeneous crowdsourcing markets. NIPS13 Workshop on Crowdsourcing: Theory, Algorithms and Applications.

[24] Difallah, D.E., Demartini, G., Cudr´e-Mauroux, P. (2013) Pick-a-crowd: tell me what you like, and I'll tell you what to do. Proceedings of the 22nd international conference on World Wide Web (WWW2013)

[25] ul Hassan, U., O'Riain, S., Curry, E.: Slua (2013). Towards semantic linking of users with actions in crowdsourcing. In: CrowdSem

[26] Bernstein (2014). Crowdsourcing and the Semantic Web (Dagstuhl Seminar 14282) . 2014. Abraham and Leimeister, Jan Marco and Noy, Natasha and Sarasua, Cristina and Simperl, Elena. In Dagstuhl Reports, Vol. 4, Issue 7 ISSN 2192-5283.