



QROWD

Quality Repair for Open Web Data through crowdsourcing

Proposal full title: Quality repair for open Web data through crowdsourcing

Proposal acronym: QROWD

Date of preparation: 23 April 2014

Work Programme Strategic Objective:
ICT 15 - 2014: Big data and Open Data Innovation and take-up

Target Outcome addressed: a.2) Innovation and technology transfer

Coordinating person: Martin Kaltenböck / E-mail: m.kaltenboeck@semantic-web.at

Tel/fax: +43 1 4021235-25 14-1915 / +43 1 4021235-22

List of participants

Participant no.	Participant organization name	Part. short name	Country
1 (coordinator)	Semantic Web Company	SWC	AT
2	BROX IT-Solutions GmbH	BROX	DE
3	Ontos AG	ONTOS	CH
4	Renewable Energy and Energy Efficiency Partnership (REEEP)	REEEP	AT
5	Unister	UNIST	DE
6	University of Leipzig, Institute of Applied Informatics	INFAI	DE
7	University of Southampton, Electronics and Computer Science	SOTON	UK

Table of contents

1. EXCELLENCE	3
1.1 OBJECTIVES	3
1.1.1 <i>Problem statement</i>	3
1.1.2 <i>The QROWD value proposition</i>	5
1.1.3 <i>Main outcomes of QROWD</i>	5
1.2 RELATION TO THE WORK PROGRAMME	6
1.2.1 <i>Focus of the project</i>	6
1.2.2 <i>Solutions and data value chain services</i>	7
1.2.3 <i>Cross-sectorial, cross-border and cross-lingual scope of the project</i>	8
1.2.4 <i>Business perspective, milestones and market validation</i>	8
1.3 CONCEPT AND APPROACH	9
1.3.1 <i>Overall concept</i>	9
1.3.2 <i>Approach and methodology</i>	10
1.3.3 <i>Positioning of the project according to technology readiness levels</i>	25
1.3.4 <i>Related activities</i>	25
1.3.5 <i>Sex and gender analysis</i>	27
1.4 AMBITION	27
2. IMPACT	31
2.1 EXPECTED IMPACTS	31
2.1.1 <i>Impact on technology development</i>	31
2.1.2 <i>Impact on the availability and market take-up of innovative tools for data quality management</i>	32
2.1.3 <i>Expected impacts listed in the work programme</i>	32
2.1.4 <i>Impact on societal challenges</i>	35
2.2 MEASURES TO MAXIMIZE IMPACT	36
2.2.1 <i>Dissemination and exploitation of results</i>	36
2.2.2 <i>Communication activities</i>	45
3. IMPLEMENTATION	47
3.1 WORK PLAN — WORK PACKAGES, DELIVERABLES AND MILESTONES	47
3.1.1 <i>Overall strategy of the work plan</i>	47
3.1.2 <i>Timing of the different WPs and their components (Gantt chart)</i>	49
3.1.3 <i>Detailed work description</i>	50
3.1.4 <i>Graphical representation of component dependencies (Pert diagram)</i>	60
3.2 MANAGEMENT STRUCTURE AND PROCEDURES	60
3.2.1 <i>QROWD structures</i>	60
3.2.2 <i>Decision procedures</i>	62
3.2.3 <i>Quality control</i>	63
3.2.4 <i>Planning and reporting</i>	63
3.2.5 <i>Risk management</i>	64
3.3 CONSORTIUM AS A WHOLE	66
3.3.1 <i>Overall consortium</i>	66
3.3.2 <i>Individual contributions</i>	66
3.4 RESOURCES TO BE COMMITTED	68
3.4.1 <i>Summary of staff effort</i>	68
3.4.2 <i>Other direct cost' items (travel, equipment, other goods and services, large research infrastructure)</i>	69
3.4.3 <i>Swiss partner</i>	69

1. Excellence

1.1 Objectives

1.1.1 Problem statement

A convincing and sustainable value proposition for open data is predicated on the availability of tools to assess and enhance the quality of such data sets. More often than not open data on the Web has been notoriously incomplete, inconsistent, or out-of-date. Without going into too much detail about the reasons for this state of affairs, what is indisputable is the fact that this problem concerns all industry sectors that have invested, or are contemplating entering this area. Consider first those organizations that have recognized the power of the open and are planning to publish some of their data assets under open-access licenses. They require benchmarks and certificates for data quality, and might be worried about the effects of third-party apps integrating their data with other, external sources of uncertain quality will have on their standing and business relationships. Then there are those organizations that make use of open data sets. They need reassurance that certain quality standards will be met by data publishers and intermediaries. Finally, let's think about providers of data-driven technologies. Their products and services have to be able to master the challenges raised by open data sets; they demand means to evaluate the quality of open data sets and repair them, which would become integral part of the data value chain technology they offer. In the end, it is a problem that affects everyone, every decision and action we take, as we use IT applications built on top of open data.

Europe's evolution towards a data-driven economy is driven and backed-up by technological progress, in particular by the transition of enterprise IT systems to data-driven Web-based and service-oriented models; the uptake of sensor networks and smart mobile devices; user-generated content; as well as open access as a principle for data publication that encourages data exchange, integration, and reuse. It is especially this last trend that motivates a growing number of public and corporate organizations across Europe, including public-administration offices, digital libraries, publishers and media broadcasters, telecom operators, or even financial institutions to expose large, valuable data sets via standardized Web technologies, most notably using linked data in combination with further related semantic technologies.

One of the basic design principles of linked data is that the management and use of data are amenable to a high level of automation. Standardized interfaces should allow applications to load data directly from the Web, resolve descriptions of unknown Web resources, and automatically integrate data sets published by different parties according to various vocabularies. However, this vision and the actual experience of consuming linked data do not yet fully match – the varying quality of a large number of data sets and of the links between them, as well as the vocabularies used (or not used)¹ challenges the development of applications built on top linked data both from a technical and a usability perspective. One of the reasons for this state of affairs is the ‘publish-first-refine-later’ philosophy promoted by the linked open data movement. While this motto served the community well in its early days, creating momentum and enabling rapid growth, it also led to data sets which were de facto not fit for use in productive environments. More than seven years later, the situation looks no different; while a small selection of popular, typically general-purpose linked data sets are curated and maintained by an enthusiastic community, many others are not only incomplete, inconsistent, or sparsely interconnected, but meanwhile hopelessly outdated. A second important aspect to consider is the very nature of the data provisioning process. The linked data lifecycle² is highly Web-centric, hence fundamentally different than what organizations might be used to from the closed enterprise environments that are characteristic to more classical forms of data management. Similarly to what was discussed earlier, the adoption of a decentralized way of publishing data had incredible effects in terms of the scale of the data made available online over the past seven years, and encouraged data reuse, exchange, and interlinking. However, it has also meant that that wealth of interconnected data sets is managed by a very heterogeneous community of data providers, with greatly diverse levels of expertise and technology know-how, and mostly underspecified data governance policies.³ This has affected not only the quality of the data they publish, but also, and perhaps more importantly, that of the global data ecosystem. The same network effects that act as a powerful multiplier when using linked data on the Web apply unfortunately as well when it comes to the fitness of use of the data ecosystem. Noise in one data set, or missing links between different data sets propagate throughout the Web of Data, and impose great challenges on the data value chain technology using the data. With Big Data conquering the IT agenda of any forward-thinking organization, making sense of the large amounts of data that are continuously

¹ <http://semantic-web-journal.net/blog/encouraging-five-star-linked-data-vocabulary-use>

² http://www.w3.org/2011/gld/wiki/GLD_Life_cycle#LOD2_Linked_Open_Data_Lifecycle

³ <http://blogs.worldbank.org/developmenttalk/open-data-is-not-enough-0>

generated by humans and the devices we use is challenging even when data is assumed to be clean.⁴ With an estimated 25% of enterprise data being flawed according to Gartner,⁵ and even less promising figures for linked and open data sets assessed in a 2014 survey by Planet Data,⁶ the current state of affairs is a serious obstacle in the path of commercial uptake and for the sustainability of existing efforts. A recent example of what could rapidly become a dangerous trend comes of all countries from the UK, considered by many as one of the pioneers of the field. Issues with data quality (DQ) are severely affecting the future of the Midata initiative of the UK government⁷, questioning the significant investment that has been made across European Member States in technology development, training, and policy making.⁸ Other sources cite similar reasons (besides unclear business models and privacy concerns)^{9,10} for the limited adoption of open data by the industry. Ultimately, a huge economic potential (think trillions of Euros, according to a McKinsey study for open, not just linked, data from 2013, see footnote 7) could be missed as many sectors struggle to produce curated, useful data sets that would reinstall the confidence of early adopters and convince less enthusiastic organizations of the social and economic benefits of the technology. As acknowledged by the present ICT-15-2014 call (among other prominent sources), open data is arguably one of the most important ingredients for the creation of powerful cross-border, cross-language, and cross-sector data products and services that are crucially needed to put Europe back on the map in the global data economy and address its most critical societal challenges.

It would be unfair to say that the community is unaware of or unwilling to change this state of the affairs. Technology suppliers, researchers, as well as data providers in several subject domains have put linked and open data quality at the top of their agendas – and their intensified efforts led to several proposals for conceptual models and frameworks, as well as a large array of quality assessment methods addressing specific quality dimensions (see Sections 1.3 and 1.4 later). As a result, we now have a better understanding of the types of quality indicators that are meaningful in the context of linked open data and their wider implications for the data management lifecycle. These indicators were used to study the quality of the several billions of RDF triples available online,¹¹ leading to a set of powerful best practices and guidelines that data publishers can take into account in order to increase the usefulness of their data sources, and the economic and societal benefits this data may potentially have through applications. The assessment and the cleansing of data sets are, however, by far less investigated. Although sustained action is taken to produce comprehensive studies capturing the current state of the art and to devise automatic quality assessment methods, their application scope remains limited and the actual curation of the data sets is something the community is still expected to solve.

Crowdsourcing seems to be a promising approach to tackle data quality. Putting aside any utilitarian considerations regarding the costs of data curation, its principles and mechanisms match very well the ethos of the open data movement, which strongly advocates ideas around transparency, accountability, and participatory ventures. As an example, the Open Knowledge Foundation has set up a welcomed initiative for crowdsourced data quality¹² which assembles a useful collection of case studies that convincingly showcase the benefits of the approach. Unfortunately, community-driven data curation projects such as OpenStreetMap¹³ are the exception rather than the norm. Beyond the dozen of successful examples open data evangelists are all too happy to promote, there are hundreds of thousands of data sets that have yet to find their community of benefactors. Just as with any other enterprise that leverages mass collaboration, setting up a data curation project driven mostly by volunteers remains an art more than an engineering exercise. Many domains lag severely behind in adopting such approaches. Many organizations, though open towards the idea, prefer solutions whose outcomes are slightly easier to predict and integrate into their existing workflows and IT infrastructure, or lack the expertise to apply crowdsourcing purposefully.

⁴ <http://theodi.org/blog/five-stages-of-data-grief>

⁵ <http://www.gartner.com/newsroom/id/501733>

⁶ <http://goo.gl/9V0tGc>

⁷ <https://www.gov.uk/government/news/the-midata-vision-of-consumer-empowerment>

⁸ <http://www.out-law.com/en/articles/2014/february/midata-initiative-may-have-stalled-due-to-poor-data-quality-says-it-consultant/>

⁹ <http://www.sciencewise-erc.org.uk/cms/assets/Uploads/130628-Open-Data-SI-paperFINAL.pdf>

¹⁰ http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

¹¹ <http://lod-cloud.net>

¹² <http://blog.okfn.org/2013/09/03/how-can-open-data-lead-to-better-data-quality/>

¹³ <http://www.openstreetmap.org>

1.1.2 The QROWD value proposition

QROWD offers a data quality solution with a clear business proposition that can be reliably and systematically used as an integral part of an primarily open-source technology stack for the realization of data value chains across industrial sectors and languages, complementing established data governance frameworks with sophisticated, robust, and scalable crowdsourcing-empowered quality assessment and repair features.

In the last decade linked data has seen an increasing adoption as a means to publish structured data over the Web for seamless exchange, integration, and reuse. With several billions of RDF triples available online, it has the potential to significantly lower the barrier of entry into data-intensive industries, in particular for SMEs, by offering technologies and best practices to easily get access to a wealth of interconnected data sets in a variety of domains. However this potential could soon be lost if the quality of a critical mass of these data sets remains as it is (see footnote 4), making a large share of the so-called Linked Open Data Cloud¹⁴ hardly usable in practice. In today's economy, whose success relies of the ability of stakeholders to build commercially viable data value chains bringing together and drawing insight from interconnected sources of data, these gloomy prospects threaten not just the significant investment Europe has made in promoting linked data standards and open principles, but also the capacity to grow and competitiveness of the organizations, public and private alike, that are using the wealth of data sets available online to improve their product and service offerings.

QROWD will provide the technology required to address this increasingly pressing problem at scale. We will realize a set of quality assessment and repair services for linked and open data that combine the performance of machine-driven computation with the in-depth domain insight and exquisite attention to detail of human intelligence to enhance the completeness, correctness, consistency, comprehensibility, objectivity, and timeliness of linked and open data in two key sectors of the European economy (energy and eCommerce). Any serious attempt to tackle data quality via crowdsourcing in productive environments will have to go beyond the enthusiasm of ad-hoc volunteering initiatives that lack a realistic participation and engagement strategy. In QROWD we will thus make use of various forms of crowdsourcing, including microtasks, gamification, and open challenges, all of which proved, individually or in combination, to be extremely successful in many technical scenarios that the linked and open data communities have been targeting since their very beginnings: data harvesting (across different languages), classification, annotation, enrichment, integration, and analysis. The envisioned quality assessment and repair services will be applied to produce training data and validate the output of state-of-the-art automatic curation techniques, capitalizing on the advantages offered by both human- and machine-driven approaches. Our solution will be designed as a set of standalone configurable components in order to maximize its uptake in a variety of industry sectors. We will showcase the newly developed functionality for a core of the Linked Open Data Cloud, and openly deploy it to allow data publishers and consumers to test and use it in order to solve their own data quality problems. To further increase its utility, we will integrate these services in a linked data technology stack¹⁵ that has been successfully validated in commercial setting as part of European research projects such as LOD2¹⁶ and GeoKnow¹⁷, as well as in industry projects in several business sectors. In addition we will integrate it in three popular linked data publishing and management tools (OntoWiki, OntosLDIW, and PoolParty), that will build the technology backbone for two vertical data value chains in the energy and eCommerce/online travel sectors.

Our proposal is motivated by the need of different types of data value chain stakeholders (data providers, consumers, and intermediaries, as well as technology suppliers) to improve the trade-off between the quality of the data assets they own or process and the resources to be invested in data curation. The data quality technology delivered by QROWD will be evaluated in terms of usability, accuracy, scalability, and costs by studying the technical and organizational demands of data curation workflows of in two key vertical sectors of the European data economy: energy management and eCommerce.

1.1.3 Main outcomes of QROWD

The main outcomes of the QROWD project will be as follows:

- Two vertical data value chains in the sectors of energy (renewable energy, climate-aware development) and eCommerce (online travel purchases) using a mix of multilingual open and privately managed data,

¹⁴ <http://linkeddata.org/>

¹⁵ <http://stack.linkeddata.org/>

¹⁶ <http://lod2.eu/>

¹⁷ <http://geoknow.eu/>

including text, structured data, and social media streams, using technology for data harvesting, interlinking, enrichment, maintenance, and analysis with a strong emphasis on data quality.

- Cross-sectorial, cross-language technology, including three popular publishing and authoring tool suites, covering the entire data management lifecycle as mentioned above, which capitalizes on crowdsourcing and a combination of statistical and knowledge-driven data inspection techniques to tackle the poor data quality.
- Configurable crowdsourced services, available as integral part of the Linked Data Stack¹⁸ and as public deployments for the community to use, leveraging the wisdom of the crowds via gamification, paid microtask, open challenges, and campaigns, to curate key assets of the Web of Data and hence enable purposeful use and maintenance.

In Section 1.3 we will elaborate on how we will measure the success of the project for each type of outcome just listed, including our take on validation in the market for eCommerce/travel and energy.

1.2 Relation to the work programme

QROWD addresses the second bullet point of part a) (Innovation Actions) in the topic ICT-15-2014 “Big data and open data innovation and take-up”. In the following we will revisit the call description quoted below and explain how QROWD will meet its objectives.

Collaborative projects focused on innovation and technology transfer in multilingual data harvesting and analytics solutions and services. The projects should have a cross-sectorial, cross-border and cross-lingual scope, and take into account the users' and societal perspectives. The driver in consortia should be a core of companies dedicated to focused activities with a clear business perspective with verifiable milestones and market validation.

1.2.1 Focus of the project

At its core, the project investigates the usage of crowdsourcing mechanisms to perform data collection, processing, and analysis tasks along the open data value chain. One focus area is multilingual data harvesting; we will use conTEXT, a tool based on existing NLP frameworks such as FOX (see WP2, especially Section 3.1), to semantically analyze multilingual text corpora (such as blogs, RSS/Atom feeds and Twitter) and visualize the results. The tool detects entities and their relations in multilingual text. We will integrate the crowdsourcing services developed in the project to gather feedback on the performance of the individual conTEXT components, e.g., named entity recognition and relation extraction. This feedback channel will be used to train the underlying NLP algorithms in order to assess and improve their precision and recall. Ontos will use the crowdsourcing-enabled conTEXT tool and integrate it with their own tools such as Eventos and the Ontos Linked Data Information Workbench (OntosLDIW). A first business scenario pursued by Ontos is concerned with information integration for CRM (Customer Relationship Management) - multiple data sources will be linked, enriched, and exposed within the CRM GUI via the newly developed functionality. A second scenario aims to improve the quality of Ontos' news aggregation service; in this case news content will be augmented with information from open data sets as well as from private data sets owned by the news provider. To be truly useful for a news and media company, such data mash-ups need to fulfil very high quality standards – the crowdsourcing-enabled data quality assessment and repair tools developed in QROWD will make this possible for several languages (English, German, Russian etc.) and at scale (by using crowdsourcing only when it is needed and using near-real-time methods, see WP1 in Section 3.1). The technology will be used in combination with OntoWiki as part of the eCommerce data value chain to produce a rich, interconnected data collection on topics relevant to traveling (see WP5 in Section 3.1).

REEEP's business and vast network of partners will benefit from multilingual data harvesting as well. This component of the project will equip them with services able to cross borders and sectors through the translation of key concepts in the areas of renewable energy, energy efficiency and climate-compatible development. Data harvesting will help enrich the Reegle Content Pool, by which portals that have implemented the Reegle Tagging API¹⁹ can have access and shared related documents enriched with machine-readable annotations. Crowdsourcing will offer an economic, scalable, and robust way to create accurate multilingual definitions of key concepts in the Reegle Thesaurus, and greatly improve the multilingual tags via the Reegle Tagging API. As REEEP's www.reegle.info Web site and API Reegle.info offer the Reegle Thesaurus definitions in English, French,

¹⁸ <http://stack.linkeddata.org>

¹⁹ <http://api.reegle.info/>

Portuguese, Spanish, and German, the data harvested will lead to more meaningful translations and ease interaction in the community. The Reegle Tagging API is based on Semantic Web Company's PoolParty Semantic Information Suite software product.²⁰ In QROWD PoolParty will be extended with advanced quality assessment and repair services thereby enabling the deployment of a better software product that is already in great demand in many ongoing SWC commercial projects, and will be offered to future customers as well.

Unister's B2C services are located in many countries and available for many languages. Equally diverse are the amounts of data available for each geo-political area, and its fitness of use. This is mainly the result of the different levels of process and technology maturity of the individual data providers. To be competitive, Unister's service proposition must increase its information coverage; however, this cannot come at the cost of the customer experience. Quality assessment and repair must be performed independently of the primary sources, on which the eCommerce provider has little control. Upon curating this wealth of data in QROWD for Unister's B2C Web portals, the applications in the different countries will have access to much more (as in, orders of magnitude more) accurate information by leveraging on the links that will be defined across language versions. Visitors' 'dwell time' will grow by more than 100% just because of this improved information offering. Most B2B services in the eCommerce market focus on a limited set of product attributed (e.g., for hotels one would expect to find information about price, facilities, rooms, and board). Richer information, ideally available in different languages, leads to novel services, and increased turnover. Moreover, the fact that such data sets will be published at high quality standards on the Web will prepare the ground for new ventures; start-ups across Europe could pick them up using standard APIs and build their own business on top.

Several analytics services are developed as part of the project. For instance, we perform an extensive analysis on the usage of linked data sets via queries. This application of data mining techniques to linked data sets has established itself as a viable means to capture external requirements to evaluate a data set continuously against the needs of data consumers (see WP2 in Section 3.1).²¹ It also offers a feasible approach for data maintenance, informing effort-conscious data publishers about the fitness of use of their data assets and signaling aspects which require their immediate attention. In addition, analyzing usage also allows updating data sets in response to evolving community best practices in terms of vocabulary use and hence lessens the efforts for integrating external data sources consistently. Finally, this type of analysis gives data providers a mechanism to quantify the success of an open data set, which is important to legitimate open data publication from an economic standpoint, and to develop realistic and convincing business models. In the context of usage mining, human intelligence is needed to prioritize and rank queries and interpret information needs, a costly task given the popularity of some endpoints. Besides query analytics, we will deliver a comprehensive suite of quality assessment and repair services, including statistical, structured, and schema-oriented quality analysis (see WP2, Section 3.1). They will illustrate the full power of microtask crowdsourcing, applied to train algorithms and validate their results for anything from automatically derived rules to schema-level mappings. These analytics components will be extensively evaluated on data sets used in the two vertical data value chains we will build for REEEP and Unister. Central to the idea of open data are the connections between data sets; links to external sources add value to one's data and diversify the ways in which it can be used. We will employ the state-of-the-art, highly scalable link discovery tool LIMES and equip it with the crowdsourcing services to perform a rich, fine-grained link analysis which goes way beyond the shallow computation of equivalences between entities, taking into account context-specific insights that are de facto impossible to encode in any automatic approach. This is also relevant for scenarios in which enterprise data is integrated with open data (of uncertain quality).

1.2.2 Solutions and data value chain services

The outcomes of the project are a set of tools and sector-independent data-centric services. First, we will provide standalone components for each form of crowdsourcing. Additionally, we will devise a method to deal with more complex workflows to optimize the results of the services in the standalone components. In particular, customized versions of TurkIT and Turkomatic for recursive tasks, built on top of clickworker and CrowdFlower will be implemented together with several services for near-real-time scenarios. Moreover, we will offer public services for the community to curate their data by providing public endpoints via the Ontos RDF store called OntoQUAD. The Linked Data Stack will be used as a repository to deploy these services. We understand data curation as a core support activity in the data management lifecycle; as such, crowdsourced data curation features will be added to existing robust and scalable services for data harvesting, inspection, integration, and analysis brought in by the partners allowing them to master the shortcomings and fully exploit the benefits of current open data sets. To

²⁰ <http://www.poolparty.biz>

²¹ Series of USEWOD workshops, see <http://people.cs.kuleuven.be/~bettina.berendt/USEWOD2014/>

facilitate technology transfer towards data value chains we will do the same for three comprehensive tool suites, OntoWiki, OntosLDIW, and PoolParty, and further develop these to meet the demands of two business cases in the energy (REEEP) and online travel purchase (Unister) domains.

1.2.3 Cross-sectorial, cross-border and cross-lingual scope of the project

QROWD's approach is by design cross-sectorial. To start with, the technology developed in the project is applicable to any scenario that is confronted with the daunting task of data curation – as discussed in Section 1.1, this affects at least 25% of all data worldwide ('Dirty Data' is a Business Problem', Gartner²²), according to some market research firms. Then, the project will use this technology to realize two high-yield data value chains in two different sectors with international reach. SWC's products are used to realize a shared Content Pool to stimulate innovation and entrepreneurship in the renewable energy, energy efficiency, and climate areas. They span across multiple sectors in terms of data origins, but also stakeholders that need to collaborate in order to develop clean energy solutions and to devise new policy frameworks to mitigate climate change and quantify and determine national-level responses to this threat and potential impacts. This is an area that affects all industry sectors of the economy, every Member State of the European Union, and the entire world – a wide range of stakeholders need to come together, share information, and come up with cross-cutting solutions to these societal and environmental challenges. Through its profile and international network, REEEP offers digital services enabling this collaboration.

The QROWD consortium is indeed cross-border as REEEP's tools and solutions break down silos of information among governments, non-governmental agencies, multilateral agencies, and corporate portals. Policy makers and governments within the European Union and the global community need access to cross-border information provided through REEEP's Content Pool, the Reegle Thesaurus and the Tagging API in order to prevent replication of effort, turn information into actionable knowledge, and make long lasting effective and efficient decisions. The multilingual facets of REEEP's information management solution, further advanced by participation in this project, will help transgress language barriers and greatly improve communication.

Unister's eCommerce business has a strong cross-lingual and cross-border component. While Unister's primary market is Germany, its B2B and B2C services are provided in different countries, most of them located in the Europe, but some also in Asia and the US. QROWD will provide the foundations for the development of better services in these markets. The cross-lingual aspect of the project will benefit the society since we will significantly improve the quality of several popular linked data sets, for example, DBpedia, LinkedGeoData, Freebase, Eurostat etc. We will add missing multilingual labels, comments as well as missing values of attributes, links within and between data sets through crowdsourcing. Overall, we will reduce errors and enhance the quality of the existing data sets. Also, QROWD will help curate data from the language level to a semantic level, and, by that, also enable cross-border interoperability at the level of organizations and other initiatives. Thus, the curated data sets will provide enhanced access to, and value generation on (public and privately originated) open data resulting in a significant amount of clean data sets, which can be purposefully used across multiple business sectors and scenarios.

1.2.4 Business perspective, milestones and market validation

QROWD has a strong industry orientation and is driven by business needs elicited in clear business and exploitation plans. Section 1.3 describes the business case around clean open data for each of the two verticals and for each commercial partner using the business model canvas paradigm²³. Section 2.2a includes detailed business plans for each of them. The consortium consists of five industry (SWC, BROX, ONTOS, REEEP, UNIST) and two research partners (INFAI and SOTON) with a successful track record in technology development and innovation (e.g., DBpedia framework, OntoWiki, LIMES, USEWOD, data.gov.uk) as well as in serving global 500 customers.. The aim of the project is on technology transfer and the realization of market-tested vertical data value chains (see Section 1.3). In addition, Ontos will integrate the new crowdsourcing-empowered curation services into the OntosLDIW that is based on the GeoKnow Generator (see WP3 in Section 3.1). We will evaluate our solution in terms of usability, accuracy, costs and scalability. Additionally, we will showcase the QROWD curation services as a core of the Linked Open Data Cloud, and openly deploy them to allow data publishers and consumers to test and use them in order to solve their own data quality problems (see WP1 in Section 3.1). For the two verticals, the market validation will take place within the customer segment linked data for enterprises and governments for open data; a list is available in Section 1.3 below.

²² <http://www.gartner.com/newsroom/id/501733>

²³ <http://www.businessmodelgeneration.com/canvas>

1.3 Concept and approach

1.3.1 Overall concept

QROWD will develop a cross-sectorial, cross-language quality assessment and repair solution for open and interconnected data sets. We will use a combination of automatic (statistical, inference-based) and crowdsourcing (microtasks, gamification, open challenges, community campaigns) features that together with data harvesting, analysis, and interlinking build a powerful technology stack that can fundamentally improve the fitness of use of the Web of Data. This technology is transferred into two verticals, realizing market-tested data value chains using text, structured data, and social media streams in several European languages. The results respond the real needs of technology suppliers, data providers, and consumers to find commercially viable answers to the increasingly pressing, costly problem of data quality.

QROWD will provide a comprehensive, open source quality assessment and repair solution that combines human and computational intelligence to enhance the completeness, correctness, consistency, comprehensibility, objectivity, and timeliness of open and linked data sets. To do so we will make use of several forms of crowdsourcing, including volunteering, microtasks, gamification, and open challenges, all of which proved, individually or in combination, to be extremely successful in many technical scenarios that the linked data community has been targeting since its very beginning: data collection, classification, annotation, as well as data integration.²⁴ The envisioned crowdsourcing-enabled functionality will be applied to produce training data and validate the output of state-of-the-art automatic curation techniques,²⁵ capitalizing on the advantages offered by both human and machine-driven approaches in terms of accuracy, costs, speed, and scalability. This will be realized as part of an ecosystem of hybrid (human + machine intelligence) services that span over the entire data management lifecycle, responding to specific curation and human interaction needs. Automatic curation will use statistical and inference-based approaches as offered by existing tools in the Linked Data Stack (RDFUnit,²⁶ CROCUS, ORE)²⁷ as part of fundamental data management activities such as data harvesting (using context), interlinking (LIMES)²⁸ and maintenance (USEWOD).²⁹

We will showcase the QROWD curation services for a core of the Linked Open Data Cloud, and openly deploy them to allow data publishers and consumers to test and use them in order to solve their own data quality problems. To further encourage uptake, the services will be integrated in three popular linked data publishing and management tools brought in to the project by the consortium partners: the OntoWiki knowledge engineering and application development platform by BROX/INFAL,³⁰ the OntosLDIW for data processing and management by ONTOS,³¹ and the PoolParty thesaurus editor by SWC.³² When used within these tool suites, crowdsourcing will serve different technical tasks related to linked and other sources of data (e.g., text, Twitter), which are incorporated in the tools. It will seamlessly support the knowledge acquisition process in OntoWiki by asking the crowd to collect and corroborate statements in a knowledge base; create and maintain links; and undertake documentation activities that are otherwise challenging to be executed automatically (such as selecting the most informative image for a given entity). Analogously, the crowds will solve conceptual modeling questions in PoolParty, both in a standalone fashion, and complementary to mining and learning techniques. Finally, we will apply the approach to semantically enrich text documents in the OntosLDIW in the context of machine translation, named entity recognition, and entity linking.

To ensure that QROWD fully matches end-user needs and expectations, we will evaluate the usefulness of our approach as part of OntoWiki, OntosLDIW and PoolParty, as these tools are applied to assist data practitioners in their daily activities in enterprise environments in two key sectors of the European economy: energy and eCommerce. We will set up lab and field experiments with data practitioners from the linked data community (10-15 subjects per study) and customers of the industrial partners (5-10 subjects per study at Unister and REEEP). In this context we will first identify the forms of crowdsourcing that are most likely to lead to the desired data quality enhancements, determine how to best set the associated parameters (in terms of reward models, interaction with

²⁴ See <http://behind-the-enemy-lines.blogspot.com/2010/10/what-tasks-are-posted-on-mechanical.html/> for a general overview, <http://sites.google.com/site/amtworkshop2010/> for examples of text-related tasks, and <http://www.gwap.com/> for media-related tasks.

²⁵ Such as the LODMiner, which applies graph metrics and machine learning to identify missing properties, see <http://lodminer.net/>.

²⁶ <http://aksw.org/Projects/RDFUnit.html>

²⁷ <http://stack.linkeddata.org>

²⁸ <https://github.com/GeoKnow/LIMES-Service>

²⁹ <http://usewod.org>

³⁰ <http://ontowiki.net/Projects/OntoWiki>

³¹ <http://www.ontos.com/>

³² <http://poolparty.biz/>

and within the crowd, spam prevention etc.), and study the trade-offs between costs, time, and expertise in the three scenarios. Data practitioners in the end organizations may serve different roles in the evaluation, depending on the form of crowdsourcing applied. In the two verticals they will assess the usability of the crowdsourcing-enabled versions of OntoWiki, OntosLDIW, and PoolParty, respectively (see below, and Section 3.1, WPs 4 and 5); in addition, they might be involved in validating the results obtained through crowdsourcing as domain experts, or become part of the crowd and contribute to the accomplishment of the technical tasks. OntosLDIW will also pursue a separate evaluation study around CRM and news management (in WP3, see Section 3.1 and business model later in this section). As a result of applying this multi-faceted evaluation strategy we will deliver a set of data curation tools that are not only non-intrusive regarding established work and social practices within an organization, but also deliver optimal outcomes in terms of accuracy and costs.

While using the tools in the two vertical scenarios (and as part of our public campaigns and challenges, see Section 2.2) we will significantly improve the quality of several popular linked data sets (e.g., DBpedia, LinkedGeoData, Freebase, Eurostat)³³ by adding missing information (labels, comments, also multilingual, values of attributes, links within and between data sets etc.), correcting such information, and selecting among (inconsistent) variants. The availability of a purposeful linked and open data curation solution will help data providers make the most of the data they own and expose, and will boost reuse, thus making an important contribution to the realization of the original Web of Data vision.

1.3.2 Approach and methodology

In this section we will further develop the concept just introduced along a number of dimensions. We will start by arguing for the business case for clean open data in Section 1.3.2.1, which is then taken up by two vertical scenarios which are fully aligned with the business strategies of the QROWD industrial partners (Sections 1.3.2.2 and 1.3.2.3). An outline of their strategy is presented using the business mode canvas paradigm in Section 1.3.2.4. Then we introduce the main components of the technology meeting these business needs, including KPIs and market validation approach in Section 1.3.2.5.

1.3.2.1 The business case for clean (open) data

Data governs nearly every part of our private, social, professional and civic life.^{34,35} The ability to capture its implications will offer enormous competitive advantage to any business³⁶ - this applies to the data produced and owned by an enterprise, but also to everyone else's, be that business partners or customers, or, increasingly, the hundreds of thousands of data sets published online using open-access licenses or other (fermium) models in data marketplaces. No matter where it comes from, this data deluge will need to be stored, analyzed, and turned into actionable knowledge³⁷ - challenges which become even more daunting when dealing with data of uncertain quality. This aspect of data management and governance, despite its scale and consequences (25% of all data is flawed according to Gartner)³⁸ is still to be solved.³⁶ In fact, the market of data quality tools is one of the fastest-growing in the enterprise software landscape³⁹. As discussed in Section 1.1, a similar, not very encouraging state of affairs prevails on the Web of Data, threatening the investment that has been made (mostly) by public administration and policy makers in pushing forward open access and transparency directives and releasing their data artifacts.

There are several business-driven reasons to deal with data curation, not only for enterprise, but also for open data. Let's start with data analytics, which is conducted to identify and understand current and upcoming business situations and inform decision processes. Carrying out (big) data analytics on top of erroneous, incomplete data can have fatal consequences.⁴⁰ Another well-known business case requiring correct data to be effective is (direct) marketing.⁴¹ It is common that enterprises acquire large customer data bases which need to be filtered to prepare

³⁴ Agrawal et al.: Challenges and Opportunities with Big Data, Computing Community Consortium (2012), URL <http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf>

³⁵ Watters, Audrey: The Age of Exabytes - TOOLS AND APPROACHES FOR MANAGING BIG DATA, ReadWriteWeb (2010), URL <http://de.slideshare.net/readwriteweb/exabytes-rww-final>

³⁶ <http://goo.gl/HA74d3>

³⁷ http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

³⁸ <http://www.gartner.com/newsroom/id/501733>

³⁹ Ted Friedman (Gartner Inc.): Magic Quadrant for Data Quality Tools. G00252509, Oct. 2013

⁴⁰ Curry, Edward and Freitas, Andre and O'Riain, Sean: The Role of Community-Driven Data Curation for Enterprises. In: Wood, David (Ed.) Linking Enterprise Data. pp 25--47, http://dx.doi.org/10.1007/978-1-4419-7665-9_2, Springer, 2010.

⁴¹ Hernández, Mauricio A. and Stolfo, Salvatore J.: Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. In: Data Mining and Knowledge Discovery, vol. 2, no. 1, pp. 9--37, <http://dx.doi.org/10.1023/A%3A1009761603038>, 1998.

marketing initiatives; such information can rapidly become outdated and crowdsourcing has proven instrumental to ensure that it remains up-to-date and that incomplete records are augmented with information available online or on social media. The public sector offers several interesting cases as well. Besides the success stories discussed in Section 1.1, another noteworthy example could be the publication of legislation acts in the UK,⁴² which can be used by citizens to document and back-up their claims.⁴³

There are a number of challenges that need to be addressed in order to implement a scalable, accurate, and affordable solution to data curation:³⁶

- Quantity and complexity - The structure and the number of records of a data set greatly influences the effort required to curate the data. For example, to validate a spreadsheet with 5 columns and 100 records is feasible for an individual. Assessing the quality of 3.2 million DBpedia instances⁴⁴ cries for alternatives.
- Update rates: The dynamicity of the data impacts the ways in which it needs to be inspected and updated. As an example, the GoodRelations vocabulary⁴⁵ is stable, but a data set with products and services described using it will likely to change frequently.
- Required efforts: The quality of the raw/input data set and the tasks to carry out for establishing a high quality data set (cf. Sect.) are a crucial factor as well. For instance, it is easier to double-check the syntactical as the semantic correctness of attribute values.
- Availability of experts: Independent from the data it-self, the amount of available experts which can execute the quality process is critical.

These aspects should emphasize the complexity of any data curation exercise; similarly complex is the question of return on the data quality investment. In the well-known setting in which a business attempts to increase productivity and profitability while minimizing cost and risks, the impact of poor-quality data could be described as follows:⁴⁶

- Increased costs - Costs incur for the detection and correction of problems in the data and if flaws in the data sets are recognized too late leading to delays and re-iterations of production processes. Additional expenses are caused if service-level-agreements are not met.
- Decreased revenues - Customer relationship management is less effective when acting on incomplete or outdated customer data, which may lead to customer attrition, but also to lost opportunities to generate revenue.
- Decreased confidence - Flawed data maybe result in organizational trust problems and further to suspicion. Incorrect data impairs decision-making and forecasting.
- Increased risks – This is a frequently occurring problem in every sector. For example, if governmental regulations or acts are missed a business may be subject to penalties. Incorrect data may result in risks for life, e. g., in healthcare if blood types are incorrect or the altimeter in aircrafts show up wrong information.

In the remainder of this section we will introduce two vertical data value chains which are confronted with acute data quality problems, as well as the approach pursued in QROWD to mitigate the negative effects just discussed.

1.3.2.2 *The energy data value chain*

This scenario brings together the commercial interest of two organizations, each with its own agenda, targets, business partners, and customers (see below). SWC supplies the technology required to build, maintain, and use vocabularies and multilingual thesauri, while REEEP offers subject matter expertise, publishes the digital assets, and offers services on top of them for the broader energy community. As such, the data value chain to be realized in QROWD is based on the business demands of REEEP as a provider of information management services in the energy domain complemented by the product development strategy of SWC's PoolParty tool suite. Figure 1.3.1 illustrates this relationship.

⁴² <http://www.legislation.gov.uk/>

⁴³ <http://blog.okfn.org/2012/10/04/worlds-first-real-commercial-open-data-curation-project/>

⁴⁴ <http://wiki.dbpedia.org/Ontology>

⁴⁵ <http://www.heppnetz.de/ontologies/goodrelations/v1.html>

⁴⁶ Loshin, David: The Data Quality Business Case: Projecting Return on Investment. Whitepaper, 2006.

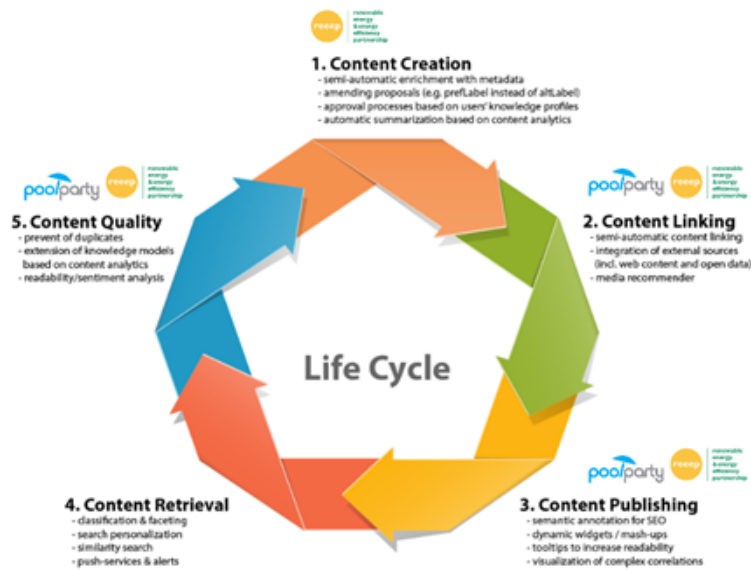


Figure 1.3.1 The energy data value chain by REEEP and SWC

REEEP’s information management offering and its stakeholders

Climate change is the home of one of the most pressing ICT challenges of present times. How do we develop tools that facilitate collaboration and information exchange across national, organizational, and cultural borders to work on one of the world’s most critical global challenge? REEEP’s www.reegle.info is an information management tool that aims to alleviate this situation. Through QROWD REEEP’s services will make the most of the information they manage by being able to master the imperfections many data sets on the Open Web suffer from today. This will benefit multiple stakeholders and users of REEEP’s product, including multi-lateral organization, government agencies, policy makers, NGOs, clean energy and climate compatible development knowledge managers, clean energy and climate compatible development experts, academic institutions, researchers, corporations, the general public, and future generations. Figure 1.3.2 lists some of the most important REEEP partners, which will be involved in the evaluation of our approach.



Figure 1.3.2 Selection of REEEP customers

www.reegle.info is used by many organizations across the world. Existing and future customers of the platform have a stake in a global asset, which we are collectively working to mitigate and adapt to, as it is threatening all our natural resources and life on the Planet. We will use three examples to illustrate the range of features supported by www.reegle.info. The information in brackets “()” refer to the steps depicted in Figure 1.3.1.

- Multilateral organizations such as the World Bank use Reegle’s Tagging API for knowledge management of the Climate Smart Planning Platform. All information across the site is consistently categorized through this tool to ease navigation, and facilitate knowledge management and sharing. The platform integrates

(Step 2, see Figure 1.3.1) trusted products and services from multiple sources and organizations covering all aspects of low carbon development and climate resilient planning into a unified catalog. Users can search, filter and facet (4) via a simple and friendly user interface built on top of this catalog. Content in the platform is very diverse and covers most economic sectors.

- The Green Growth Knowledge Platform uses the Reegle Tagging API as well. In an initial step they tagged over 600 green growth resources for their eLibrary, while keeping open the option to use the tool in order to push content back to the Reegle Content Pool. The main benefit of using the Reegle's information management product is that it allows content managers to efficiently tag, categorize, and organize (Step 1) the growing body of green growth data and knowledge published on the platform.
- The Climate Technology Centre and Network (CTCN) use the Reegle Tagging API and its thesaurus to provide users with the most relevant (Step 5) news, policies, and programs to support the development of national strategies for priority climate adaptation or mitigation technologies. CTCN is unique in its role of facilitating access to climate technology-related information. It covers sources from all over the world and its services have a similar global reach.

Going beyond this small selection of examples, it is ultimately the end-users searching these portals (governments, NGOs, policy makers, managers in the private sector interested in green technologies) who reap the most benefits from REEEP's information management tools. By improving the quality of the data used to organize and tag Reegle's Content Pool it is most likely that REEEP will be in advantageous position to gain new customers for its tools as well. They will mostly be owners of online portals that will contribute information to the Content Pool, using the Reegle Tagging API to analyze and categorize this information. The Content Pool will be leveraged to generate on-demand reports for paying customers with specific requirements that can only be satisfied through advanced analytics on top of the content. This exploitation model is becoming increasingly important, as more and more organizations feel the pressure to react to IPCC's recent findings on the climate change.

PoolParty as technology enabler of the energy data value chain

The REEEP data value chain focuses on the management and curation of vocabularies and multilingual thesauri, as well as on data integration from various sources on top of those thesauri. To do so they use using SWC' PoolParty authoring environment. Through QROWD PoolParty will be extended into advanced data curation features. Thesaurus managers will have the option to execute quality checks such as:

- Label conflicts (pairs of concepts that are labelled identically) → Fewer duplicate concepts.
- Incomplete language coverage (i.e., concepts lacking documentation in a language that should be supported) → Better performance in translation usage scenarios.
- Cyclic hierarchical relations (Most often considered illogical and are probably mistakes) → Better understand ability.
- Orphan concepts (i.e., concepts not connected semantically to other concepts in the thesaurus) → Find concepts that remain unused.
- Missing out-links (that are, concepts not connected to third-party resources on the Web) → Provide additional information without duplication.
- Broken links (as in, references to other resources on the Web that provide no data) → Maintain a highly informative thesaurus despite the ever-changing nature of the Web.
- Full support of the semi-formally defined integrity constraints of the SKOS reference document → Compatibility with other SKOS vocabularies.

The quality assessment and repair services developed in QROWD will have a wide-ranging, positive impact on the functionality of several other components in PoolParty, thereby improving the overall quality of the Reegle Content Pool. These components are concerned with:

- Automated data harvesting (as in the REEEP scenario e.g., statistical data on energy consumption or energy production from several 3rd party sources such as UN Data, DBpedia, Eurostat);
- Data processing and integration (enrichment, interlinking, and fusion of data from several internal and external sources); and Data analysis (generation of reports with a specific purpose).

This will help REEEP to provide more precise and comprehensive data analytics for the clean energy market, which could be used, for instance, to generate timely and correct clean energy country profiles and reports. Such reports are an essential decision-making tool for REEEP customers; consider, for example, a report on the current

situation and the outlook of wind energy in Tunisia and its importance for European investors, thereby for governments, as well as project managers, and technology providers in this area.

These tasks will leverage different forms of crowdsourcing and automatic quality assessment techniques, as explained in Section 1.3.2.5 and in WP4 in Section 3.1.

Data sets relevant for the data value chain

We will use DBpedia,⁴⁷ Freebase,⁴⁸ Eurostat, UN Data,⁴⁹ and WorldBank statistical data,⁵⁰ but also data from GTZ⁵¹ and REN21⁵² as well as thesauri such as GEMET of EEA,⁵³ FAO's AGROVOC,⁵⁴ and the GBPN terminology.⁵⁵ The data is available as linked data (RDF Datacube), as well as CSV and XLS. Sometimes it is collected via Web APIS or just scrapped from Web sites. Thesauri are encoded as RDF SKOS or as MediaWiki articles or spreadsheets. Data covers different subject domains (energy efficiency, renewable energy and climate change development) in several European languages (English, French, Spanish, German, and Portuguese).

1.3.2.3 The eCommerce data value chain

One of the greatest challenges for eCommerce service providers today is to adjust their business cases to a data-driven world. Major Internet players like Google, Amazon or Apple increase their market share to very high levels and reduce the opportunities for existing, smaller players and start-ups. Due to its close-to-monopolistic market share, Google, for instance, has gained a doorkeeper position for most Internet users when it comes to accessing and finding information about virtually anything. They channel Internet traffic and shape it by their own rules. Many eCommerce companies tend to reduce their market role to a mere selling point, hence limiting their opportunities for a better turnover and profit. This is especially true for European eCommerce players, which are directly threatened by this development.

One of the reasons for this discouraging situation is related to the ability the European eCommerce sector to capitalize on the data economy. Internet giants, commonly based elsewhere, have all the means they need to collect and curate vast amounts of data, analyze it, act upon it, and therefore create a good user experience. To keep pace and match these offerings, European eCommerce providers need to innovate in terms of technology, or make better use of open data assets. In the eCommerce domain such data sets are available for most parts of the industry, and their value could significantly grow through integration.

Creating a purposeful Web of Data for eCommerce is not a trivial task. First, there is the scope of the individual data sets, which makes automatic interlinking methods hardly applicable. Most B2B players in this space operate on data sources with a very narrow topical focus (e.g., customer opinions about products, hotels located in a given area, regional prices). Although these data sets are typically of high-quality (due to their relatively manageable size, curating them manually is still feasible), they do not match well the long tail of the information needs of their customers. Their augmentation through the integration with other data sets proves difficult to solve with existing tools, as the narrow scope of the data sets is not represented well by generic technology. It is a classic case of crowdsourcing application, as humans are much better (and a much more economic source of information) at dealing with such fine-grained contextual aspects. Second, there is the question of trustworthiness and timeliness of the different data sets. If the customer has any doubts about the data exposed via the portal, she will move her quest to a competitor. Such negative effects could be created through integration with external data sets, whose quality is uncertain, or by failing to update the data when it changes. Solutions for near-real-time maintenance, realized as a combination of automatic and crowdsourced services, could offer temporary relief; however, the quality of openly available data sets, which could be used (at least in theory) in a wide range of eCommerce applications, is a problem that commercial companies are unwilling to solve. In a time when worries about the business exploitation of the hundreds of thousands of open data sets out there grow bigger, convincing the private sector, for instance, in eCommerce of their added value will have to have radically different solutions to data curation.

Unister's eCommerce offering for online travel

⁴⁷ <http://dbpedia.org/>

⁴⁸ <http://www.freebase.com/>

⁴⁹ <http://data.un.org/>

⁵⁰ <http://data.worldbank.org/>

⁵¹ <https://www.giz.de/>

⁵² <http://www.ren21.net/>

⁵³ <http://www.eionet.europa.eu/gemet/>

⁵⁴ <http://aims.fao.org/standards/agrovoc/about>

⁵⁵ <http://www.gbpn.org/databases-tools/glossary>

Unister is one of Europe's leading eCommerce technology companies. It manages more than 40 online portals worldwide, adding up to more than 10 million users monthly. They provide Internet solutions for multiple eCommerce verticals, including travel, automotive, news and media, and retail. Many popular sites in these verticals, in Germany, but also elsewhere, use Unister technology. The data value chain pursued in QROWD will focus on online travels (<http://www.unister-travel.de/>), capitalizing on sites such as <http://www.ab-in-den-urlaub.de/> or <http://www.fluege.de>, which are market leaders on their segment.

The following examples should shed light on the technology and services offered by Unister:

- Large companies exchange data sets about the products they are providing, typically using standard APIs (B2B). For example, the largest hotel aggregators worldwide trivago.com and hotelscombined.com provide access to their region tree and booking API for other Web portals to use them and open up a source of revenue. Such services are also exploited by businesses whose direct focus is not hotel booking. For example, tripadvisor.com gives the customer the possibility to book a hotel alongside reviews. Offering such services for others to invoke is predicated by the availability of high-quality data.
- New business models within the B2C verticals are mostly about high quality data. There is a wealth of data sets available online for others to use. Such information could improve the services offered by Web portal providers, leading to an increase in turnover and revenue. A representative example here is zaptravel.com - they became very popular by expanding the scope of their site strictly from hotel bookings to framing a vacation in a broader theme like beach or culture. Defining such connections manually in house is pricey and does not scale well; using an external source or data set, if available at a fair level of quality, would be a much better choice and would create opportunities for a whole new market of such focused applications to unfold.
- Managing information related to online traveling is a challenge business; travel planning is affected by many factors (geo-political, weather, traffic, construction sites etc.) and the quality of service that an eCommerce player in this sector can offer is directly influenced by the timeliness and trustworthiness of the information it provides to the visitors of the site. Keeping pace with these updates and managing the complexity of the integration exercise, with covers not just multiple domains, but also different languages demand for scalable quality assessment solutions that not only identify update needs, but are able to fetch the relevant information and add it to the data set. As we will see later on, a combination of human and machine intelligence (via crowdsourcing, to cap the costs of manual labor) is in most cases the only viable way to achieve this.

These examples are indicative of a more general problem, which appears in any eCommerce vertical. In QROWD we will develop a data value chain for eCommerce in the online travel purchase domain using advanced data harvesting, management, and analytics technology by Unister, Ontos, BROX, and INF AI, which, as part of the project, will be fundamentally expanded into sophisticated and scalable data curation features. The resulting data value chain will allow defining service level agreements on quality and time for different types of data sets, including multilingual data. This will help the data-driven enterprises, in particular eCommerce and eBusiness players, to evolve and improve their business models. In particular, extending the B2B and B2C services of Unister will not only help their own business, but also the businesses of the companies connected to these services. Additionally, it is expected that richer data sets with a high quality will also inspire entrepreneurs all over Europe to establish new ventures as the barrier to market entries will be significantly lower and confidence in open data sets will improve.

OntosMiner, Eventos, and OntosLDIW as technology enablers for the eCommerce data value chain

The eCommerce data value chain is all about an interconnected, rich information space that can be used in innovative ways to meet information needs and improve customer experience on the site. Text mining technology is providing opportunities to extract implicit knowledge and match it to entities from the Web of Data. Knowledge bases describe properties of such entities and their relationships and enable new facts to be derived through inference. Using a combination of Ontos and INF AI technology (OntosMiner, Eventos, OntosLDIW in conjunction with conTEXT) the data architect of the eCommerce provider is able to:

- Semantically annotate eCommerce products and create multilingual metadata that will help to promote the eCommerce products to search engines.
- Analysing text in multiple languages, detecting trends that can be used to create a better forecast.
- Extract named entities from natural language text that will allow to interlink them with own data sets and further link them with other open data sets. This will augment the knowledge allowing the eCommerce company to produce better recommendations to the consumers.

OntoWiki as technology enablers for the eCommerce data value chain

BROX does have long standing experience with using linked data for the integration of distributed enterprise information systems. One of the major challenges for eCommerce data is the assessment of the quality of the data harvested in the corporate as well as the continuous improvement of the quality of the data integrated from external sources. Unfortunately, data sets in enterprises usually do not have comprehensive quality indicators attached, thus their quality is often unknown. Solving these issues is at the core of the QROWD project.

A crowdsourcing-enabled version of OntoWiki will be used to manage and curate Unister’s customer data bases for the vertical analysed in the project. BROX also provides the eccenca Linked Data Suite, which in addition to OntoWiki also contains components for integrating and managing data sets in the enterprise domain. BROX will integrate the components developed in QROWD into the Unister infrastructure.

Data sets relevant for the data value chain

We will use Eurostat, GeoNames,⁵⁶ LinkedGeoData,⁵⁷ OpenStreetMap, and Natural Earth.⁵⁸ We will also exploit several closed-source data sets, in particular data feeds collecting hotels, airport, and flight connection information, CRM data, focus groups, as well as textual descriptions about domain entities. There are several data suppliers for the same category of data items. The format of the data sets are diverse, e.g., RDF representations, CSV files, JSON object s (via Web APIs), dump files of databases, text files or shape files for geo-spatial representations, as well as binary APIs on top of SPARQL and SQL endpoints. Most data sets have English labels or descriptions, some are only available with German annotations. In addition several data feeds focusing on one entity class only (e.g., hotels) provide their information in English, German, French, Italian, Spanish and Chinese

1.3.2.4 Business models of QROWD industry partners

SWC				
Key partners Domain expert groups Taxonomy publishers / providers Partners (integrating the service into their applications)	Key activities Quality validation and quality improvement module for PoolParty and as a SaaS. Procedures and techniques for experts involvement into quality assurance for taxonomies Semi-automatic checks and validation for consistence and standard conformance	Value propositions Quality approved thesauri fit for commercial sale Quality mechanisms integrated in current workflows Quality enhancement with semi-automatic procedures Involvement of expert groups into quality assurance mechanisms for the domain - thus less burdens for the core taxonomy management	Customer relationships Licensed plugin for PoolParty Suite License for stand-alone module to be integrated in various stacks	Customer segments Information- driven companies which act as a hub for domain- or theme specific taxonomies, vocabularies etc. Corporate companies with a disperse structure, several branches Companies which act as agency for trustworthy data for third parties (financial, health, pharmaceuticals, environment) Governmental and non-governmental institutions
	Key resources Quality standards for taxonomies (QSKOS etc.) Domain experts and taxonomists Open taxonomies for comparison Access to central taxonomies for several domains		Channels Product suite Existing network of sales partners Existing network of integration partners	
Cost structure 20% Hardware and maintenance 30% Rewarding crowd resources 30% Data analysis 20% Tool development			Revenue streams 60% licensing within own product suite 40% integration in third party products	

⁵⁶ <http://www.geonames.org/>

⁵⁷ <http://linkedgeodata.org/>

⁵⁸ <http://www.naturalearthdata.com/>

BROX				
Key partners The eccenca Linked Data Suite is based on open source tools that are jointly developed with the AKSW research group at University of Leipzig	Key activities IT consulting Data integration using Linked Data Knowledge management Data marketing Automotive supply chains Product development eccenca Linked Data Suite eccenca Enterprise Search Suite	Value propositions Data integration and knowledge management Publishing of linked data	Customer relationships Conferences Key account management Direct sales	Customer segments Medium to large enterprises
	Key resources Access to linked open data sources		Channels outreach at Leipzig Semantic Days existing customer relations	
Cost structure 10% Hardware and maintenance 20% Rewarding crowd resources 40% Data analysis 30% Tool development		Revenue streams 50% licensing the eccenca Linked Data Suite 50% consulting services		

ONTOS				
Key partners Research labs and universities working on the Linked Data Stack Service providers (e.g., system integrators) Software suppliers (e.g., establish OEM approach with leading ERP, CRM, BPM suppliers)	Key activities Technology: best of breath for a holistic linked data lifecycle combined with ease of use Sales/marketing: SPIN selling, key account management, partner model, best demo and customer feedback loop.	Value propositions Seamless, holistic ease of use information integration and data publishing Effortless integration into existing infrastructure - improve cost reduction and TCO Connect data silos and make new knowledge available within other apps (e.g., CRM, Web portals) or publish them as open data	Customer relationships Conventions and conferences Web demos, on premises demos Large enterprises: key account management Medium enterprises: Direct sales and partner eco system	Customer segments Enterprise (Medium to large) addressing information integration or data silos challenges Government (for linked data publishing and data integration) System integrators, enabling them deliver services to their customers
	Key resources Robust linked data technology (e.g., third party components) IP rights, IP protection Access to customer databases (verified contacts, decision makers) Skilled R&D people in semantic technologies		Channels Sales force “B2B”, conferences Web sales “B2B”, Web site promotion Partner ecosystem Indirect System integrators / Partner ecosystem Web promotion, viral marketing Academia/MOOCs	
Cost structure Employees/R&D Server hosting “Cloud” Marketing and promotion on the Web 24 h support		Revenue streams Software license model (on premise), maintenance/support Software subscription (Software-as-a-Service “Cloud”) <ul style="list-style-type: none"> - Free limited basic account e.g., storage, transactions or for Universities - Premium pro account e.g., high availability, 24 h support Consulting Partner royalty model		

REEEP				
Key partners SWC New funders such as governments, multi-lateral organizations, international organizations Open and linked data network Knowledge management experts in the renewable energy, energy efficiency and climate compatible development fields	Key activities Implement quality assurance mechanisms for the Reegle Content Pool and feedback loops for the Reegle Tagging API Implement out of the box CMS integrations for Reegle Tagging API and pushing content into the Content Pool Enhance multilingual API services for quality check and appropriate tagging in multiple languages further development and expansion of the Reegle Thesaurus	Value propositions Increased data reuse; cross boarder and cross language interoperability	Customer relationships Free service provider for improved free version of Reegle services Open and linked data consultants with customized solutions for on demand paying customers Open and linked data consultancy Openness, shared learning, knowledge management, interoperability, development project partnerships	Customer segments 60 international organisations registered and use the Reegle Tagging API Public and private International organizations working in development field such as governments and non-government organizations, multi-lateral organizations, Academics Think tanks General users – in 2013, 2 million users visited www.Reegle.info
	Key resources Demand for data to solve global problems on renewable energy, energy efficiency, and climate change		Channels Existing network International conferences on clean and renewable energy	
Cost structure Free version Social Entrepreneurship value is priceless by way of assisting European (and international) organizations in sharing and managing information and data in the renewable energy, energy efficiency and climate compatible development fields.			Revenue streams Reporting and consultation services for tailored solutions outside the free version will be quoted on demand Funding streams for improvement, expansion and quality assurance are consistent through donations from governments after initial release	

UNISTER				
Key partners Crowd, users Data supplier Partners (integrating the service into their applications)	Key activities SaaS to query properties of resources of the supported domains Accurate data sets created through fusion from other sources	Value propositions Convenience Trust Accessibility High quality data Low price	Customer relationships Pay per use (1.) Pay per download (2.) Reward loyalty and high quality answers (crowd community)	Customer segments Data-driven companies: recent data, e.g., online booking agencies Product-driven companies: distribution channel for their product or an enrichment for their data, e.g., hotel operators Information-driven companies: access to trustworthy information about a specific topic, e.g., market research, consulting
	Key resources Data suppliers Tools for (crowd-based) quality assurance Servers Crowd		Channels Own Web site Partners	
Cost structure 20% Hardware and maintenance 30% rewarding crowd resources 30% data analyses 20% tool development			Revenue streams 50% pay per use 50% pay per download	

1.3.2.5 Approach and methodology

The linked data lifecycle depicted on the left hand side of Figure 1.3.3 is the established framework for the publication and management of open interconnected data sets on the Web. Within QROWD we aim at upgrading the role of data quality assessment and repair in this lifecycle from a single activity towards a broader support activity that is concerned with the outcomes produced by all other, rather development-oriented, activities of the overall process. As illustrated on the right hand side of Figure 1.3.3, QROWD will not be just about the quality of a data set, but also about the quality of the results produced through the execution of data publication and management activities such as interlinking, classification, enrichment, maintenance, extraction, and so on.

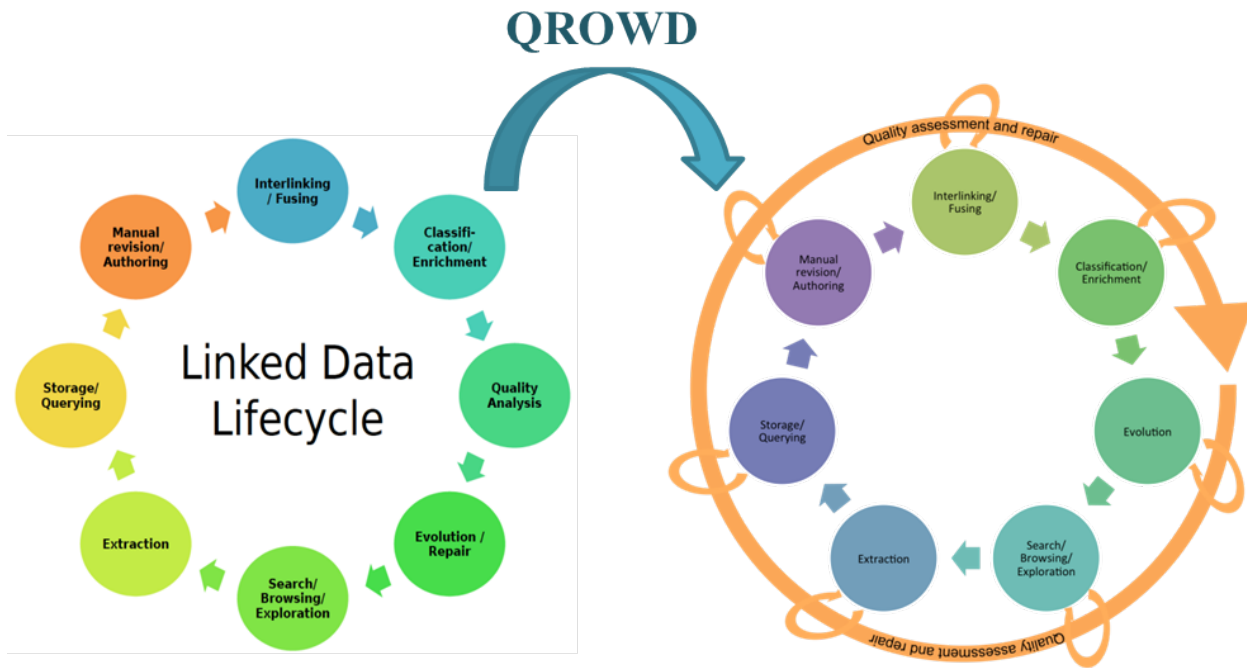


Figure 1.3.3 Quality-minded linked data lifecycle

The QROWD data quality model

Following Accenture⁵⁹ the main data quality requirements for decision-making are: (i) accuracy: data must be sufficiently accurate to avoid material distortion of the model output; (ii) completeness: databases must provide comprehensive information for the undertaking; and (iii) appropriateness: data should not contain biases, which make it, unfit for its intended purpose. We created a manual mapping between the data quality dimensions described by state-of-the-art data quality frameworks^{60 61} and the three Accenture business demands. Depending on the framework chosen, the mapping covered around 15 to 20 quality dimensions. We then selected a subset of six dimensions (completeness, correctness, consistency, comprehensibility, objectivity, and timeliness), which are amenable to crowdsourcing. Table 1.3.1 defines each of the dimensions. Some of them are also subject of automatic or semi-automatic quality assessment algorithms. The decision whether to include or discard a specific quality dimension was based on the experience of the partners in running different kinds of crowdsourcing projects and on the current understanding in the field (see also the examples in Section 1.1 as well as footnote 13). Revisiting the pertinence of this framework according to the needs of linked data consumers will be part of the project.

⁵⁹ <http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Data-Quality-Key-Solvency-Requirements.pdf>

⁶⁰ Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.* 12, 4 (March 1996), 5-33.

⁶¹ Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Commun. ACM* 45, 4 (April 2002), 211-218. DOI=10.1145/505248.506010 <http://doi.acm.org/10.1145/505248.506010>

Table 1.3.1 Data quality dimensions addressed in QROWD

Data quality dimension	Description
Completeness	Is data missing for the current tasks?
Correctness	Is the data true and free of errors?
Comprehensibility	Are the languages, symbols, units, and definitions of the data appropriate for the current consumer?
Consistency	Are there any inconsistencies in the data?
Objectivity	Is the data impartial?
Timeliness	How up-to-date is the data?

Crowdsourcing quality assessment and repair

Our crowdsourcing solution will be designed as a set of standalone configurable services in order to maximize their uptake in a variety of industry sectors and application scenarios. The configuration parameters depend on the type of crowdsourcing to be followed, and on the actual task that will be outsourced to the crowd. In the project we will cover a number of such data management tasks, and in particular those which are known to be critical to linked data consumption, but remain heavily reliant on human input despite the continuous efforts the linked data community to approach them through (full) automation.⁶² These tasks include: the creation and maintenance of vocabularies and data records, ontology alignment and data interlinking (for same-as and related SKOS predicates), as well as semantic annotation and machine translation. In a broader data-value-chain context, we will use similar methods in relation to data collection and harvesting tasks that are mostly NLP-centric and, in this project, are executed on text, tabular and social media (stream) data. The purely task-related configuration parameters will cover the type of task to be tackled, as well as inputs, outputs, and external sources to be taken into account.⁶³ The other configuration parameters of the components to be developed concern strictly to the way the actual crowdsourcing process works; for instance, for microtasks, one would specify the number of answers required for each Human Intelligence Tasks (or HITs, the unit of work in Amazon’s Mechanical Turk,⁶⁴ one of the most popular platforms in this area), the payment model, the qualification of the workers targeted, and the time the tasks will remain open on the platform; more advanced settings may also specify the way to assess the accuracy of crowd output, as well as more refined means to assign work to the crowd (see also WP1 in Section 3.1).

Each quality dimension can be studied in the context of multiple activities of the linked data management lifecycle. There are two fundamental approaches to pursue this. In a first instance one would resort to data practitioners to identify and classify issues according to the seven dimensions we selected. In a second step, one would scale the experiment to larger crowds but finer-granular tasks (microtasks) using Mechanical Turk or social networks. We have successfully tested this approach on DBpedia.⁶⁵ In addition, we will define the parameters of a crowdsourcing service depending on the actual dimension. For instance, something like comprehensibility could be tackled via exam-like questions testing if the participant has understood the meaning of a data attribute. Correctness would be addressed via a two-staged workflow in which the crowd first identifies potential sources of error and then fixes them, a pattern which is called Find-Fix-Verify in the crowdsourcing literature.⁶⁶ Table 1.3.2 gives a brief outline of the way in which each dimension would be addressed via microtasks (executed through gamified actions on social networks or as part of paid microtask platforms). As noted earlier, other forms of crowdsourcing, e.g., open challenges could be used together with microtasks to further refine the definition of the data quality problem. In any case, the final results - that is, the curated data - will be validated through a mix of manual (sample-based), and automatic techniques (voting, probabilistic reasoning etc.). They will be published according to a purpose built vocabulary as linked data in order to enable their reuse in the data management lifecycle.

Hybrid services

The Linked Data Stack comprises tools that can be jointly used to support linked data lifecycle. Table 1.3.2 below lists them, with components that will be further developed in QROWD highlighted in bold.

⁶² See also Simperl, E., Norton, B., Vrandečić, D. ‘Crowdsourcing tasks in Linked Data management’. Proceedings of the COLD workshop co-located with the ISWC2011, for a more elaborated discussion on the selection of tasks amenable to crowdsourcing.

⁶³ See footnote 5; the paper provides further details about the parameters, which could be encoded as SPARQL patterns.

⁶⁴ <https://www.mturk.com>

⁶⁵ "Crowdsourcing Linked Data Quality Assessment" Maribel Acosta, Amrapali J. Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, Jens Lehmann, In: 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia, Springer, 2013

⁶⁶ <http://projects.csail.mit.edu/soylent/>

Table 1.3.2: Data value chain technology developed in QROWD

Objective	Components
Extraction and loading	Sparqlify, TripleGeo, Virtuoso Sponger, DBpedia Spotlight, PoolParty , D2RQ, Stanbol, VALIANT
Interlinking and fusing	LIMES
Classification and enrichment	GeoLift, DL-Learner
Quality analysis	ORE: Ontology Repair and Enrichment
Evolution and repair	LODGrefine
Search, browsing and exploration	Facete, sig.ma, CubeViz, Spatial Semantic Browser
Storage and querying	Vituroso RDF Store, SparQLed, Sparqlproxy, OntoQUAD
Manual revision and authoring	OntoWiki , RDFAuthor, PoolParty

We already mentioned that QROWD develops crowdsourcing services to support data quality assessment and repair and that the project turns the quality analysis step into a parallel support activity of the entire lifecycle. Hence, the highlighted components of the Linked Data Stack will be extended by crowdsourcing support. Additionally, in order to increase the impact of the developed solutions, QROWD will also contribute completely new components to the Linked Data Stack. In particular these components are:

- **conTEXT** for multilingual harvesting fitting into the extraction and loading objective;
- **RDFUnit** and **CROCUS** for data inspection purposes enriching the set of available quality analysis tools;
- the **USEWOD** tool chain for linked data usage analysis supporting multiple lifecycle steps concerned with data set maintenance including interlinking and fusing, quality analysis, and evolution and repair.

Integration of hybrid data curation services to OntoWiki, OntosLDIW and PoolParty

The data curation services will be evaluated as integral part of the three tool suites that will enable the vertical data value chains of the business partners. To realize the energy scenario, we will extend PoolParty with services for taxonomy development, maintenance and interlinking. A list of lower-level tasks a thesaurus manager is in charge of was presented in Section 1.3.2.2. The thesaurus manager will be assisted by quality assurance and repair features that will review the taxonomy according to the six quality dimensions introduced earlier, which could be tackled using different crowdsourcing approaches as explained earlier, as well as by tools such as USEWOD and CROCUS. Besides using professional microtask platforms, we will add a series of non-intrusive quizzes for PoolParty users (or even the end-users of Reegle.info) to collect the required information. Typical quizzes could cover, for example, testing (i) if alternative labels used in the thesaurus are correct; (ii) if concepts are connected through specific types of links (skos:exactMatch, owl:sameAs); (iii) if an entity belongs to a certain class; and (iv) if an attribute of an entity has the expected value. Several variants of such quizzes will implemented, including multiple choices for any of the elements included in the underlying statements (e.g., choosing which of the target concepts matches a fixed source concept vs choosing which type of link/property best describes the relationship between two given concepts).

For the eCommerce data value chain, data curation will be undertaken within an professional publishing and authoring editing environment, which will contain automatic assessment and repair algorithms and crowdsourcing interfaces to Mechanical Turk, CrowdFlower⁶⁷ and clickworker,⁶⁸ but also to enterprise-internal resources. To leverage this second type of collective intelligence, we will look into methods to target microtask-like requests to specific individuals of groups within a (closed, thus known) pool of human labor based on their expertise, and reward contributions through gamification elements (see also Section 3.1, WP1 for a summary of the task assignment methods by which this could be achieved). The editor will include a generic component for supporting such elements, including profiling, simple input validation, scoring and bonuses, as well as badges and ranked lists. Data cleansing and repairing will cover different types of information and tasks: adding or maintaining (multilingual) products labels; creating comprehensive product descriptions; updating specific features; collecting relevant images; classifying records according a vocabulary; as well as data interlinking within the newly created data hub and to the rest of the LOD Cloud. Both, the metadata editor and the resulting catalog, will be built on top of existing linked data publishing and management technology available in the project, in particular OntoWiki. Data harvesting and annotation will be supported through tools such as conTEXT and OntosMiner/OntosLDIW.

⁶⁷ <http://crowdfower.com/>

⁶⁸ <http://www.clickworker.com>

These tools will be extended with a direct interface to several microtask platforms in order to deal with the multilingual character of the data. Typical tasks will include entity classification, labelling, and entity interlinking.

Market validation of QROWD's data curation technology

We will follow an established approach that is widely used by the industry to test new products and services. The approach is centered around a series of interviews on a representative sample of potential customers. The interviews are used to test a concept against a potential target market, leading to adjustments in design and implementation, but also to a refinement of the marketing strategy in terms of product placement and market.

Market validation in QROWD will be performed in two iterations and will concern three areas (where the third one is the most important):

- The data curation services combining crowdsourcing and automatic algorithms (see earlier as well as WPs 1 and 2 in Section 3.1)
- The extended versions of the three tool suites (OntoWiki, OntosLDIW, PoolParty, that are the result of WP3, see Section 3.1)
- The vertical data value chains developed together with REEEP and Unister, respectively.

The phases of the market validation are as follows:

- Identification and specification of the target market(s) and high-yield contacts
- Development and set up of the market validation questionnaire
- Comprehensive face to face (or online room) interviews in the target market(s); in both iterations with at least 5 but up to 10 participants per sector
- Analysis of the results and implications for product/service development and marketing strategy.

To identify the target markets the following approach will be taken:

- First, and probably most obvious, is to talk to people or companies directly in the respective target market, means to speak to (potential) target customers. There are many ways to find out who is in a respective target market. On the one hand the members of the QROWD consortium has already a broad customer base and the target markets are thereby already well defines , as the 3 ICT vendors can use the existing customer base to present and demonstrate their extended products and services and thereby generate a helpful feedback cycle on top of existing customers. Also the business case partners are well established in their respective target markets and can make use of these existing contacts to get in touch with the potential customers of the new products and services. But above this also social networks and social media are great tools to be used for this identification process as for instance LinkedIn or Twitter beside others. Understanding the potential touch points between the product & services and the customer process is key.
- Then we will find experts who target the same market as the QROWD partners do, but are not competing with them. Such experts may sell a different product or service but are targeting the same market. For example the ICT products of the 3 QROWD ICT vendor partners are often integrated into 3rd party information management systems (e.g., PoolParty of SWC has integrations in Atlassian Confluence as well as in Microsoft SharePoint in place). The vendors of these software products in which the QROWD tools are integrated know the respective market very well but do NOT act as direct competitors and thereby can be very helpful in market validation. Also included in this group are analysts and well-respected domain experts. Such expert groups can be extended additionally by evaluating the relevant authors / users of social networks or blogs et al in the respective field of interest.

In parallel to the identification of the target market (stakeholders) the market validation questionnaire is being created and then used for the series of interviews that needs to be accomplished with the representatives of the respective target markets – such an interview takes around 45-80 minutes usually and is a set of product presentation and discussion and the answering of a structured questionnaire. The whole interview follows a structured process from the beginning until the end to ensure that all questions are covered completely as well as that a comparison is possible in the course of the result analysis.

In all cases at least the following questions should be raised and answered:

- Whether the product solves problems that are pervasive in the target market
- Whether or not target companies would buy it / make use of it (and why / why not)
- What the overall workday process of the representatives of the target markets is

After the management of the face to face or online room interviews the gathered data will be analysed. Analyzing the data is where it gets interesting. Then the existing knowledge of the QROWD partners in respect to their own markets and products & services is important to efficiently and correctly interpreting the data. The answers will be compared and conclusions drawn. At the end a high-level summary for management that includes key findings and suggested changes in product positioning, value statement, and target market will be created.

The following REEEP partners will be involved in the market validation interviews for the establishment of the energy data value chain: World Bank (Climate Smart Planning Platform partners will validate data and provide contacts to UN Data department), UN-REDD, UNEP, UN's CTCN and their data driven network, GGKP, NREL, Eldis, The German Government, The Technical University of Denmark, weADAPT, Energypedia, Energy and Environment Finland, Open Energy Information (EI), Climate Tech Wiki, FAO Stats.

The market validation of the online travel data value chain will use three prominent German sites (<http://ab-in-den-urlaub.de/>, <http://fluege.de/> and <http://reisen.de/>) as well as three European sites in the UK, France and The Netherlands (<http://travel24.co.uk/>, <http://vol24.fr/>, and <http://vlucht24.nl/>)

1.3.2.6 Summary of outcomes and performance indicators

Outcome	Performance indicator	Baseline	Target
Energy data value chain	Demographic and geographic analysis of organisations using the service New customer acquisition /number of organisations using the services Customer feedback on the quality of services	Lots of NGOs and international organisations worldwide (mainly developing countries) Around 50 organisations using the data and information services of the Reegle Tagging API and the Content Pool Small number of testimonials & feedback by customers so far (~7) Visibility of REEEP as information- and data service provider in the fields of clean energy and climate change development	More outreach in EU and to governments and also first outreach to relevant industries. 100+ organisations using the service – NGOs, NPOs, governmental institutions but also industry High number (20+) of high quality feedback on the services High visibility for REEEP as a provider of powerful information- and data services in the domain of clean energy and climate change development
eCommerce data value chain	Geographic analysis of organisations using services Number of language-specific services New customer acquisition Higher turnover/revenue	Services are available in German, French and English only, not available with specific versions for small countries (e.g., BeNeLux, south east european countries) Existing big & strong customer base but still enough outreach in current regions for new customers	Services can be brought stronger to new regional markets like Eastern or South Eastern Europe by higher multilingual data quality assurance Win new customers by providing higher quality information and data and thereby services Raise turnover with QROWD results ~2-3%.
OntoWiki	New customer acquisition More downloads of OntoWiki More publications about OntoWiki Lower customer attrition	Lots of users in academia ~ 1.700 downloads on GitHub Around 40 publications in place	Attract more industry users and governments to make use of OntoWiki Reach 3.000+ downloads More high quality publications and accepted papers at relevant conferences
OntosLDIW	Better customer retention New customer acquisition Lower customer attrition Geographical analysis of	Satisfied customer base Customers in Russia, UK and Switzerland Low number of customer loss	More satisfied customer base because of better products (data quality management) More customers EU-wide but

	potential customer / markets Higher turnover / revenue Better market position	Switzerland, Russia and UK (Europe) Ontos is visible in the mentioned (regional) markets	also broaden the customer base in Russia No customer loss Reach new target markets (geographically) Raise turnover and revenue ~15%. Raise visibility as a provider of information management solutions.
PoolParty	Better customer retention New customer acquisition Lower customer attrition Geographical analysis of potential customer / markets Higher turnover / revenue Better market position	SWC has a satisfied customer base Customers all across Europe, in Australia and US Very low number of customer loss EU, US and Australia in place SWC is visible in the mentioned (regional) markets as software vendor of enterprise semantic information management solutions (mainly master data & enterprise taxonomy management)	Reach a more satisfied customer base because of better products & services in data quality. More customers in existing markets and entry of new regional markets Low customer loss rate Reach new target markets as Eastern Europe, better coverage of US market but also Asia. Raise turnover ~10-15% on top of the QROWD results in PoolParty. Raise the visibility as a key software vendor for semantic information management and Linked (Open) Data solutions worldwide.
Data curation services combining crowdsourced and automatic methods	High use of the components – direct use and integration	Not available yet	Continuously growing number of use and integration
Public endpoints and service deployments	High number of users High visibility in relevant communities	Available without integrated quality mechanisms Already heavily in use but somehow not satisfying the users	Higher usage of the endpoints Mainly more positive feedback on the quality of the services
Crowdsourcing services	High use of the services – direct use and integration in applications High visibility in relevant communities	Not available yet	Continuously growing number of use and integration Visibility through publications and presentations at events etc
Crowdsourcing vocabulary	High use / reuse of the vocabulary in applications High visibility in relevant communities	Not available yet	Continuously growing number of use and integration Visibility through publications and presentations at events etc

1.3.3 Positioning of the project according to technology readiness levels

Our assessment of the TRL is based on each individual partner’s experience from research and from industry, commercial related projects.

Table 1.3.3 TRLs of QROWD technologies

Technologies	Positioning of Technology Readiness Levels for QROWD																	
	Current TRL									TRL after QROWD								
	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
Crowdsourcing						6											7	
Microtasks						6											7	
Gamification							7											8
RDFUnit					5												7	
DL-Learner					5												7	
Link Discovery					5												7	
ML Data Harvesting					5													8
NLP Frameworks						6												8
LOD Usage Analysis				4													7	
OntoWiki						6											7	
OntosLDIW					5												7	
PoolParty							7											8

Technologies around crowdsourcing, microtasks, and gamification have gained momentum in various business domains in recent years,⁶⁹ but are rarely used in an operationalized manner for data quality assurance, be that linked, open, or enterprise data. QROWD aims to transfer these technologies into established data governance frameworks, and hence raise their TRL substantially. Technologies related to WP2 have been heavily used in research (and some industry) projects. They will be augmented with robust, but lightweight services using advanced crowdsourcing and multilingualism features. NLP frameworks including open source, commercial solutions, and tools developed by the research community within publicly funded projects, have been on the market for some time. We will enhance them with an integrated human feedback loop in order to increase precision and recall (also in WP2). OntoWiki and OntosLDIW are both tools that have their origin in EU-funded projects, e. g., LOD2 and GeoKnow; though stable, they have not been widely deployed in productive environments. Both will be improved through services developed in WP1 and WP2 in order to reach the next level of maturity (WP3). PoolParty has been used in numerous commercial projects. Through QROWD it will finally obtain the fine-grained and scalable data quality assessment and repair mechanisms taxonomists have been long crying for (WP4).

1.3.4 Related activities

The following table gives a comprehensive overview of related projects, data sets and software products, which are related to activities and aimed outputs of QROWD:

Name	Description	Relation
EU funded projects		
LOD2	Aims to develop infrastructure technology and best practices that fill the chasm between structured-linked-data and applied model logic & reasoning.	Reuse the quality assessment tools, practices and results in QROWD along the data set DBpedia and OntoWiki
DIACHRON	Takes on the challenges of evolution, archiving, annotation and data quality in the context of LOD.	Reuse the tools, metrics and results of the quality assessment tasks
GeoKnow	Aims at bringing geospatial knowledge integration to the Linked Data Web.	Adapt tools and techniques for spatial vector data quality assessment, reuse measures in a linking framework, which can then be crowdsourced to assess the quality.

⁶⁹ <http://www.gartner.com/technology/research/gamification/>

BIG Data Public Private Forum	Works towards necessary efforts in terms of research and innovation, as well as for technology adoption and supporting actions for the successful implementation of the Big Data economy.	QROWD complements the BIG project, and contributes specific, cost-effective means for quality assessment and repair to the data value chain.
INSEMTIVES	Produced methodologies and tools for the massive creation and feasible management of semantic content.	Crowdsourcing methods and services in QROWD will built on by the findings of INSEMTIVES.
Other projects		
DoW	Swiss CTI funded project on the orchestration and visualization of Linked Data processes.	Crowd-based services, gamification and conTEXT will be used to enhance the DoW functionality
Open Fridge	This energy efficiency genome project (FFG, national) analyzes real-time data of refrigerators to develop improved strategies for energy saving.	Processes of quality assurance may be attached to a given crowd along with feedback mechanisms, which help comment and classify data.
CTCN Reegle Thesaurus	Develops the Reegle Thesaurus in the areas of climate change adaptation, particularly the economics of and technologies to mitigate or adapt to climate change.	Drupal Plug-in will increase accessibility to the Reegle Tagging API, showcasing the ongoing support and development of the Reegle tools received to ensure all funding is consistently built upon with further development.
Data sets		
DBpedia	One of the largest data sets in the LOD Cloud, extracted from Wikipedia.	DBpedia contains multilingual data. Widely used, poses several quality challenges that reduce returns
LinkedGeoData	Core data set for spatial open data.	Crowdsourcing as a means to expand and curate well-researched quality issues
Software projects		
Linked Data Stack	Integrated set of linked data management tools	Will be used as technical backbone of all software components in QROWD.
RDFUnit	Test-driven data debugging framework	Will be used in combination with human computation
ORE (Ontology Repair and Enrichment)	Quality assessment tool for schemas; knowledge engineers improve OWL ontologies by fixing inconsistencies.	Will be used in combination with RDFUnit and human computation to create fully-fledged quality assessment and repair services
TripleCheckMate	Tool for crowdsourcing quality assessment of linked data (expert crowds, contests)	Will be configured to include any data set relevant in QROWD.
LIMES	Large-scale link discovery framework for the Web of Data.	Crowdsourcing to assess and improve precision and recall of link specifications and validate LIMES outputs.
OntoWiki	Tool for agile, distributed knowledge engineering scenarios.	QROWD will support the knowledge acquisition process in OntoWiki by asking the crowd to collect and corroborate statements in a knowledge base; create and maintain links; and undertake documentation activities.
PoolParty Suite	Supports scenarios relevant for enterprise metadata and information management, includes UnifiedViews, PowerTagging, and Semantic integrator along with qSKOS (SKOS quality checker).	Reuse qSKOS for automated quality checks on controlled vocabularies, reuse the Semantic Suite for thesaurus management, text mining and data integration tasks, UnifiedViews will act as a LOD management suite to schedule and monitor ETL jobs.
eccenca Linked Data Suite	Collection of tools for linked data management and integration (developed by BROX, part of the Linked Data Stack).	OntoWiki is part of eLDS and will be extended in QROWD, the extensions will become part of eLDS.
OntosLDIW	Integrated framework for linked data	Will be enhanced with the crowdsourced

	management	services.
OntoQUAD	RDF store	Will be used to store the RDF data and endpoints.
OntoDix	Ontos' tool for RDF data authoring and visualization	Will be integrated to the OntosLDIW and therefore also connected with crowdsourced services enabling the repair and quality of data and links.

1.3.5 Sex and gender analysis

As stated in the Toolkit Gender in EU-funded research⁷⁰ (2009, Directorate-General for Research, EUR 23857 EN) gender-sensitive research takes a twin approach: it encourages an inclusive approach and it integrates gender and diversity into the research content throughout from the initial research idea to dissemination and the transfer of results into productive environments. Following this approach QROWD will implement the measures and policies listed in the table below.

Topic	Measure or policy
Gender balance in the project consortium	Whenever new staff is hired for the project all partners are bind to tender the posts in a gender-aware manner and to give always applications of women advantage, in case they fit equal or better than an application by a male candidate. With the very best credentials – the University of Southampton, for example, has held the Athena SWAN award since 2006 - ⁷¹ we will naturally apply the gender and diversity policies that all consortium members have already established.
Working conditions allow all members of staff to combine work and family life	Most activities in the project are supported by cloud-based technology. This ensures that physical presence in offices, labs, and computing facilities is not required at all times.
Manage and monitor gender equality aspects	Gender equality aspects (e.g., implementation of the measures and policies described here) are subject to every meeting of the General Assembly (monitor) and the project management board (manage).
Explicitly and comprehensively explain how gender issues will be handled	The project management board will initiate a briefing for all parties of the consortium, where gender equality aspects for the day-to-day-work will be discussed and measures arranged.
Institutions, and journals that focus on gender included among the target groups for dissemination along with mainstream media	We will enlarge our media and partner outreach with a special focus on media and partners with positive programs towards gender and diversity. Efforts and results will be documented in the minutes of the General Assembly.

1.4 Ambition

QROWD will push forward the state of the art in applied crowdsourcing, crowdsourcing technology, and data curation (see Table 1.4.1). As an innovation project enabling the realization of vertical data value chains, it will fundamentally enhance key components of a technology stack that has been funded through a number of European projects and found industrial adopters in several industry sectors (see Table 1.4.2).

Table 1.4.1 Advancing quality assessment in QROWD

Area	Status	Innovation
Data quality assessment and repair	Manual, expensive	Affordable quality assessment processes at predictable costs and target quality levels
		Higher planning flexibility as quality assessment is partially outsourced, and it is easier re-allocate funds in peak times rather than hiring new personnel
		For each different type of task, we will configure the corresponding services according to inputs, outputs, as well as

⁷⁰ http://www.yellowwindow.be/genderinresearch/index_downloads.html

⁷¹ http://www.southampton.ac.uk/diversity/gender/athena_swan.page

		budgetary, time, and quality constraints. Automated data cleansing, reduction of duplicates, repairing broken data links to produce a new and better Web of Data
Crowdsourced quality assessment and repair	Several encouraging success stories in different industry domains, but for very specific tasks No operational framework, little to no integration to data governance or data authoring and management environments	Set of robust deployments of existing tools for data quality assessment and repair augmented with crowdsourcing services, accessible via Web APIs
		Standalone, configurable components for many popular forms of crowdsourcing, can be assembled into more complex data processing workflows
		Optimization methods to handle complex workflows, real-time constraints, and large-scale tasks
		Crowdsourcing translations and definitions of key terms statistically analyzed and optimal data results automatically published in tools for public user interface
Continuous data harvesting to ensure data timeliness	Manual, expensive, resource intensive, snap shot	Affordable and focused automated data harvesting, scraping and cleansing based on tools and feedback from crowdsourced interfaces (gamification, user engagement, microtasks) leading to better statistical trend analysis and higher quality of on-demand reports generated through data analytics.

Table 1.4.2 Advancing products and services in QROWD

Product	Purpose	Advancement
Crowdsourcing & data quality core components		
RDFUnit + ORE	RDFUnit is a tool used to assess the quality of instance data. ORE works at the schema level.	We will connect RDFUnit and ORE to provide crowdsourcing services in order to reduce the effort required by the maintainer to perform these two tasks.
USEWOD tool chain	USEWOD analyzes SPARQL query logs down to the level of atomic triples.	We will turn the tool into a Web application and deploy an instance of the service to be used by linked data providers. Open data sets will be monitored by the community in real-time. Maintenance activities for such open assets can be hence better planned and executed by an expert crowd, possibly in combination with other flavors of decentralized problem solving.
LIMES, SILK and KnoFuss integration with conTEXT	These are link discovery frameworks for Web data that detect similar entities and perform duplication elimination. Link specifications can be created manually or via machine-learning with human feedback. conTEXT semantically analyzes text corpora and visualizes the results.	We will implement a human feedback loop based on microtask crowdsourcing to generate data to train the underlying NLP algorithms. By exploring and reporting on the state-of-the-art link discovery tools this task will also set the course for a gold standard data set and benchmark in this space, which will be developed as a community initiative.
Enhanced data management products		
Crowdsourcing enabled OntoWiki, OntosLDIW and PoolParty	These tools are mainly used for linked data authoring, named entity extraction, as well as data management and integration.	We will deliver new releases of three tool suites empowered with different types of crowdsourcing-based data quality assessment and repair functionality, usable to create clean-data value chains.

SWC - PoolParty		
PoolParty Thesaurus Server	Core component of the PoolParty Semantic Suite to model SKOS thesauri and enterprise taxonomy	The Thesaurus Server will be enhanced by crowdsourcing and data quality components and also integrated with the existing qSKOS quality service. Thereby the issue of quality management in thesaurus modelling is optimised as well as the quality management of the linking of different knowledge models and thereby the enrichment of a Thesaurus.
PoolParty Semantic Integrator	The full PoolParty Semantic Suite that enables data and information integration from various internal and Web-based sources to realise integrated views and search based applications as enterprise services.	The PoolParty Integrator will be enhanced in the area of quality management along the whole data value chain: data harvesting, data enrichment and processing, data storage and data analysis and querying and thereby also in data visualisation.
ONTOS - OntosLDIW		
OntosMiner and Eventos	Ontos products for text processing and information extraction.	Crowdsourcing will improve thesauri, co-referencing functionality thus leading to a better F-measure quality for the analyzed and extracted named entities.
OntosLDIW	Integrated linked data management framework by Ontos	The outcome will be an extended workbench with more substantial data quality support.
OntoQUAD	RDF store by Ontos	The store will be tested and improved for scalability and performance to ensure the smooth operation of the new data quality mechanisms.
REEEP energy data value chain		
Reegle Thesaurus	The Reegle Thesaurus created via SWC's PoolParty is the source of the concepts used by the Reegle Tagging API and the Reegle Glossary (English, French, German, Portuguese, and Spanish). The thesaurus models concepts using hierarchies and relationships. Thematic areas covered include clean energy, UN-REDD, green growth, climate compatible development, and climate change.	REEEP will expand the multilingual definitions of each concept in the thesaurus using accurate and affordable crowdsourcing techniques. Users will be asked to define thesaurus concepts via a game interface. A statistical analysis of the user contributions will identify the most common translation. Alternatively, we use paid microtasks to rank the different translations. The same technology will be used for concept labels and further documentation.
Reegle Tagging API	The Reegle Tagging API extracts concepts from text and tags documents using the Reegle Thesaurus. The aim is to ensure consistent terminology and tags, in order to accurately categorize, connect, and articulate content on the Web to the end-user.	The algorithm used in the Reegle Tagging API will be enhanced by modelling results more closely to the hierarchy and relations of concepts within the Reegle Thesaurus, thereby displaying concept scores in a more robust way which represents both frequency and relevance more accurately. Quality assurance and continuous improvement mechanisms will be implemented into the Reegle Tagging API sending reports of terms to REEEP which frequently come up in documents but are not currently in the Reegle Thesaurus. This information will come with statistical analysis showing the source of the information being analyzed frequently suggesting concepts not in the thesaurus, as well as the frequency such documents are tagged, and the frequency such concepts are present in the text. REEEP can take action and approach organizations for further development of the Reegle Thesaurus to provide continuous improvement of the Reegle Tagging API.

<p>Reegle Content Pool</p>	<p>The Reegle Content Pool contains content from several sources about clean energy and climate change development. Based on the concepts a document is tagged with the Content Pool service recommends similar and related digital artifacts.</p>	<p>Automated quality checks for duplicate or broken links to content, repairing them automatically and implementing notifications for the content provider to replace with new link. Trend analysis of the crowd outputs to identify frequently tagged concepts for on-demand reporting; automated harvesting makes implementing the Reegle Tagging API on stakeholder sites much easier, reducing costs of resources and efforts to implement the Reegle Tagging API on more sites, crossing more borders and sectors through the tools.</p>
<p>UNISTER eCommerce data value chain</p>		
<p>Unister B2C travel portals</p>	<p>These portals offer tourism services such as hotel and flight booking to end customers. Currently forced to be very focused in their choice of service due to challenges to secure access to high-quality sources (see Section 1.3.2)</p>	<p>The outcome will be an improved Web portal with a stronger focus on information gathering. Existing information will be augmented, e.g., by travel motives, information about focus groups, current events. A user is likely to spend much more time on the site while getting himself informed (dwell +100%). Because of increased user experience this will lead to a stronger position of Unister travel Web portals.</p>
<p>Unister B2B services</p>	<p>Unister's B2B services currently focus on a) many domain-specific product offer services, the current prices for e.g., a flight connection on a specific date; and b) on data analysis – aggregated user opinion about e.g., a specific hotel. The aim is to make such services reusable to scale up the business.</p>	<p>The number of services will increase dramatically. However, the focus will be on data supply. Unister's goal is to develop better such services for all industry sectors in which it has a presence, leading to a substantial increase in revenue. An increase of 500% is realistic, even more since no such services are available on the market. Services should rely on standard vocabularies to enable the reuse by partners and competitors. This will lead to a higher market penetration as Unister will be active as early mover.</p>
<p>BROX - eccenca Linked Data Suite</p>		
<p>eccenca Linked Data Suite</p>	<p>The eccenca Linked Data Suite (eLDS) is a collection of tools for integrating and managing linked data</p>	<p>OntoWiki, as part of eLDS, will be improved with crowdsourcing capabilities. eLDS will also be extended with a component for quality assessment.</p>

2. Impact

2.1 Expected impacts

Data-centric technology is one of the most rapidly developing areas in ICT, as the ability to make sense of large amounts of data becomes business-critical, assuming a place, alongside labor and capital, as an essential factor in production (McKinsey Global Institute, May 2011).⁷² This trend involves both privately owned enterprise data (master data, product, customer, and market data), but also openly available data sets, typically published by public administration agencies and other open access and transparency advocates. The economic potential of open data has continuously evolved over the past years, and, with an estimated direct impact on the EU27 economy at €32B and an expected annual growth rate of 7% (The Open Data Economy, CapGemini Consulting⁷³), it is said by many to be one of the next game changers of the digital age, alongside smart wearable devices, nanotechnologies, and other futuristically inspired innovations.

One of the reasons for these promising prospects is its impact on a wide range of information products. In combination with enterprise data or other closed-sourced data sets, it can give a company an inexpensive means to gain competitive edge by enabling new, richer forms of analytics, better information provisioning, and creating large service ecosystems. The McKinsey Global Institute states in their report “Open data: Unlocking innovation and performance with liquid information” of October 2013⁷⁴: “Open data—machine-readable information, particularly government data, that’s made available to others—has generated a great deal of excitement around the world for its potential to empower citizens, change how government works, and improve the delivery of public services.” Besides its social benefits, open data is expected to generate significant economic value; according to the McKinsey report just cited, seven industrial sectors alone (transportation, healthcare, education, energy management and a few others) could create more than \$3 trillion a year in additional value as a result of open data. Encouraged by such forecasts, we have witnessed the rise of open-data-centric entrepreneurial businesses (in their hundreds in Europe)⁷⁵ as well as the release of novel and greatly improved information services unlocking this high-yield resource.

To create global impact and enable substantial growth, consuming open data needs to become a commodity. Commercial uptake is slowly advancing, however, investment in open data management and integration is often perceived – and, as noted in Section 1.1, there is some evidence to support this perception – as being too high compared to the expected returns. One of the reasons for this state of affairs is, perhaps unsurprisingly, data quality, a challenge which has haunted data governance managers since the beginnings of ICT. Consequently, the market of data quality tools is steadily expanding - Gartner⁷⁶ predicts a CAGR of 16% by 2017 to nearly \$2 billion in software revenue. “Across the landscape of enterprise software, this market is among the fastest-growing”, they note. Such data quality tools and services are important along the whole data value chain – from harvesting to processing, storage, analysis, and publishing. Some would argue it is essential for the further adoption of open data sets, given the significant public investment that was made in policy development and releasing core data assets, and the excellent forecasts of market research authorities.

QROWD will operate in this very space. It will deliver a comprehensive, mature solution to curate open, interconnected data sets; the resulting data authoring and management technology will impact the overall data economy, across industry sectors and applications. In the following we will discuss how the outcomes of the project will affect the data management market, as well as the expected impact of QROWD along the H2020 work programme.

2.1.1 Impact on technology development

QROWD will have an impact on technology development at four different levels as follows:

- **Data:** QROWD will stimulate the emerging market of open and linked data by building bridges between several areas of technology development whose integration has remained largely unexplored: quality assurance, standards, automated multilingual data harvesting, and interlinking.

⁷² http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

⁷³ <http://www.capgemini-consulting.com/the-open-data-economy-0>

⁷⁴ http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

⁷⁵ For instance, the Open Data Institute in the UK, one of the most active incubators in the area, supports ten start-ups so far <http://theodi.org/start-ups>

⁷⁶ Ted Friedman (Gartner Inc.): Magic Quadrant for Data Quality Tools. G00252509, Oct. 2013

- **Methods and mechanisms:** QROWD will produce new, sophisticated methods for data quality assessment and repair which will be published as open-access software and deployed for the community to explore and use. They are validated in the area of linked and open data, but could easily be applied to other technical scenarios developers and data owners.
- **Integration to commercial software:** As a first step in the innovation life cycle, we will transfer the technology into commercial data authoring and management tool suites, which can be adapted to other domains and vertical applications by third parties avoiding the steep costs of new implementations. The Linked Data Stack of technologies will be further developed and extended into advanced features for data curation based on a purposeful combination of crowdsourcing and automatic methods.
- **Vertical data value chains:** The technology will be taken up to build two vertical data value chains, with high exploitation potential from an economic and societal point of view. In the process, the data-centric technology underlying these value chains will be substantially closer to market readiness.

All in all, QROWD lowers the barriers for developing and applying bespoke services for data quality assurance, which can particularly encumber start-ups and SMEs focusing on downstream markets of the data value chain.

To ensure that the tools resulting from the project will be truly useful to developers we will reuse existing, widely-accepted standards, format, architectures, and technologies. In addition, the outcomes of QROWD will feed into existing standardization initiatives, in particular within W3C, following up on ongoing contributions by members of the consortium (see Section 2.2).

2.1.2 *Impact on the availability and market take-up of innovative tools for data quality management*

QROWD will produce several software artefacts that are by design open and reusable across different sectors and application scenarios, unlocking open data innovation for the industry represented in the consortium, and for the broader data economy. We will realize:

- New approaches implemented as generic (open-source) components for data quality management in the form of crowdsourcing, including microtasks, gamification, open challenges, and volunteering (SOTON & INFAL)
- New releases of market-established enterprise data authoring and management tool suites, now offering a comprehensive solution to data curation (SWC, BROX, Ontos)
- Usage of these tool suites to improve the productivity and potential impact of vertical data value chains in the areas of clean energy and eCommerce-based travel management (REEEP & UNISTER)

These innovative software tools will be delivered to the market through the industry partners along their exploitation strategy (see also Section 2.2.) to be made available for the broader European data economy. QROWD outcomes will be taken up into standardization groups within OMG and W3C. We will also provide a mix of open source software and, where appropriate, intellectual property protection for the tools developed within the project and seek to seed commercial spin-outs and licensing deals based upon these innovative tools.

2.1.3 *Expected impacts listed in the work programme*

The potential areas of impact of QROWD are many and varied. In the following we link these areas to the expected impacts listed in the work programme.

Enhanced access to & value generation on (public & private sector) open data resulting in hundreds of multilingual applications reusing tens of billions of open data records used by millions of EU citizens.

Considering the emerging data market the main problem hindering efficient open data use is often the quality of the data. Although Web data and open data has already a high quality standard when published we know from several activities in the field of open data business (e.g., the Open Data Business Day in Vienna, Austria in 2012⁷⁷ or the continuous work of the Cooperation OGD Austria⁷⁸) that industry is seeking for better quality of open data sets to be integrated into their respective data warehouse to enable enhanced data driven applications and services. Considering the business to business (B2B) market the data consumer will only agree to pay for the data if the data provider can provide a respective SLA (and this is NOT the case in open data or Web data harvesting). Altogether, we can conclude that data quality is the major blocker for the market entry of a (open and linked) data set.

⁷⁷ <http://ogdb.eventbrite.com/>

⁷⁸ <http://www.data.gv.at/hintergrund-infos/cooperation-ogd-austria/>

From an economic perspective, solving the data quality problem will lead to substantial economic growth by opening the vast number of data-driven industrial applications powered by the similarly vast number of available data sources. Opening data-driven industrial applications to publish vast amounts of new data in standardized formats will lead to foster the reuse of such data in innovative ways by application developers, essentially leading to millions of European citizens having usable interfaces to this data impacting their daily lives.

In stating that “Data is the new oil”, Commissioner Neelie Kroes, already in 2012⁷⁹ established an analogy which is still useful today. For example, this analogy holds for quality: Raw or unprocessed crude oil is not generally useful in industrial applications. Oil was not usable for many industrial applications without the first step: refinement. Thus, a main economic impact is providing an approach for a cheap and high-performance process for refining data regarding a set of data quality dimensions. The development of such a process and its corresponding technology stack is the paramount goal of the QROWD project. QROWD will provide unique methods, tools, and processes for achieving these requirements; in particular the multilingual aspect will be considered by choosing linked data as underlying data representation technology.

We are tackling the main economic challenges obstructing high quality data: (1) lack of human resources in particular in the field of European SMEs (data workers, scientists and data architects); (2) lack of performance; and (3) high costs in particular for multilingual applications.

These three main economic challenges are further articulated as follows:

1. Currently, many human resources are required to check the data quality. Our goal at QROWD is to extend the existing tools with capabilities of achieving high quality with little or no required input from the comparatively scarce and expensive resource knowledge that are knowledge workers. Data quality assurance processes are typically based on logical and statistical approaches. However, they are only capable of achieving a quality level of around 70-90%. The technological impact of QROWD will be derived from the combination of these existing technologies with crowd-sourcing approaches. Through the new concepts we strive for achieving a data quality of 99% (f1-measure). From an economical point of view, this will be cutting the requirements for data workers for this task to a minimum as they are pricy and only available in small numbers. From the technological point of view, we deliver the key concepts and software to replace the knowledge of the engineers by the swarm intelligence of many but untrained people.
2. Current data quality processes lack performance, as they need interaction by knowledge engineers as well as intensive computations. QROWD will tackle both sides. Our goal is to decrease the required time of data workers by a magnitude of 10 and to increase the computing time by a magnitude of 2. Overall this will increase the performance by a magnitude of 5 which will lead to the economic impact that a high-quality data set can be updated in short intervals, e. g., daily instead of weekly, allowing the application to provide accurate, up-to-date information via open data in many time-to-market-driven industrial fields such as news, stock markets, the health sector as well as clean energy development driven by European industries such as the Renewable Energy, Energy Efficiency and Climate Compatible Development fields.
3. Europe is fostering cross-border services and thereby cross-lingual services needs to be created and established. Multilingual applications have enormous high costs in development and continuous maintenance. QROWD works on this problem mainly by making use of multilingual Thesauri (enterprise taxonomies) in the respective applications that enable the implementation of more cost efficient multilingual applications by using innovative Language Technology. QROWD supports this movement by the integration of multilingual data quality mechanisms into the authoring tools of the three software vendors in the consortium.

Reducing the costs of accurate, standardized reliable data is addressed by QROWD in two ways. First, our processes will significantly advance the state of the art by combining less costly, yet purposeful human interaction with high-performance algorithms. Therefore, the costs for a quality assurance process will be cut down to 20%, cf. (1) and (2). Second, in order to be successful internationally, applications must be currently ported to several languages through manual translation. This is a very time consuming and costly part of the application publication readiness and hinders application use in smaller countries if the required language was not ported to. Supported by the power of linked data, the QROWD technology stack will ease the translation of the underlying data sets to arbitrary languages by lowering the costs up to 25% as in traditional approaches, cf. (3).

⁷⁹ http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm

Viable cross-border, cross-lingual and cross-sector data supply chains involving hundreds of European actors in a robust and growing ecosystem capable of generating sizable revenues for all the actors involved and SMEs in particular.

In traditional supply chains, such as in the automotive industry, the role of most suppliers is not that of a raw material supplier, but instead they build goods from parts they receive from other suppliers. Their business model is based neither on providing raw materials or a finished product, but instead on refining and combining other goods. An individual supplier does not have to deal with the complexity of the final product, but instead the overall complexity of the end product is broken down into manageable pieces. In the emerging data-driven economy, most data suppliers still fulfil the role of raw material suppliers. Instead of plugging into an existing data supply chain, each new data provider has to manage the entire complexity of the data sets it produces. Due to the growing size and complexity of data sets, this incurs high costs and may lead to error. On the other hand, similar to traditional supply chains, a data supply chain distributes the effort for building a data set by allowing gradual improvements. A precondition for viable data supply chains is interoperability between borders and languages as well as a high data quality. The methods and technologies developed at QROWD will ease the implementation of data supply chains through the following measures:

As stated above, QROWD will reduce the costs of assessing and repairing the quality of data sets. The quality assurance (QA) components will be based on reusable QA mechanisms and services that lower costs of rolling out data-driven services in different markets with different languages. The corresponding repair components make use of crowdsourcing to reduce the data curation costs drastically.

QROWD will provide a comprehensive stack for assessing and improving the quality of data sets. The project partners will devise a framework covering well-established quality dimensions and state-of-the-art automatic algorithms complemented by human-computation capabilities by which different forms of crowdsourcing will be applied to preserve and improve the data quality along specific dimensions. Through the service-based nature of the technology stack, specific tools could be easily integrated in different stages but also platforms of a data supply chain facilitating various tasks of the linked data management cycle.

QROWD will enable cross-border interoperability between data suppliers due to two approaches. First, the use of linked data for representing multilingual data sets allows QROWD to abstract from the language level to a semantic level boosting the cross-border interoperability. Second, different kinds of curation tasks are specially made for crowdsourcing, e. g., the collection of human readable labels in different languages, which will be available as a generic service.

Tens of business-ready innovative data analytics solutions deployed by European companies in global markets.

Proper data analytic methods and tools, integrating organizational domain specific data with open data, are the foundation to get insights and, thus, indicators to shape businesses or research directions, from the growing amount of data. Examples stem from various domains, e. g., retail, where Target Corporation⁸⁰ is able to predict pregnancy of its customers⁸¹, or insurance, where it is critical to detect insurance frauds⁸². Carrying out such data analytics requires sophisticated tools but high quality of data as well. Otherwise, erroneous data will lead to misunderstanding and finally to wrong conclusions. For that reason, data quality management steps including assessment, repair, and cleansing have to be realized up-front in the analysis^{83,84}. A study executed by Trifacta⁸⁵, a start-up in the area of data cleansing, state that “Flawed analyses due to dirty data are estimated to cost billions of dollars each year. Discovering and correcting data quality issues can also be costly.”

For these reasons, the results of the QROWD project have three impacts to foster sophisticated data analytics.

During the project runtime we will quality repair a set of widely-used linked open data sets, e. g., DBpedia or LinkedGeoData, which leads to benefit for all (analytical) applications based on them. Furthermore, the project partners enhance the quality of their enterprise data sets, e. g. at www.Reegle.info, during and beyond project runtime which results in better services for their customers relying on the complete data value chain.

⁸⁰ <http://www.target.com>

⁸¹ http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=1&

⁸² http://www.researchgate.net/publication/257034952_Managing_Data_Quality_With_ERP_Systems_-_Insights_From_The_Insurance_Sector/file/60b7d524401f9b02fa.pdf

⁸³ <http://dx.doi.org/10.1016/j.ijinfomgt.2014.02.002>

⁸⁴ <http://www.sciencemag.org/content/331/6018/703.full>

⁸⁵ http://www.trifacta.com/wp-content/uploads/2014/01/Trifacta_DataTransformValue_WP.pdf

QROWD will supply a set of methods and open source tools to support the quality repair of (linked) data. These solutions could be adopted by a broad range of companies and non-/governmental institutions to enhance the data quality before publishing (role of data provider) or analyzing (role of data user) the data.

There are multiple use cases for employing the results of QROWD beyond the original concept of improving the data quality. For example, a great achievement of the project would be the foundation of a spin-off or a portfolio extension of project partners targeting the data analytics market since the new tools allow for controlling assumptions or to identify logical contradictions in linked (open) data sets.

Availability of deployable educational material for data scientists and data workers and thousands of European data professionals trained in state-of-the-art data analytics technologies and capable of (co)operating in cross-border, cross-lingual and cross-sector European data supply chains.

Although the main focus of QROWD is not the publication of educational material for data scientists and data workers we do expect that QROWD can provide an important benefit to all of these data professionals by innovating the data quality management approach for data scientists and data workers. Publications and reports on data quality mechanisms and crowdsourcing services will be published under an open license to be available for use and reuse and thereby can easily be integrated into curricula of relevant training courses or similar like toe School of Data or the Coordination and Support Action of ICT 15 Big data and open data innovation and take-up on capacity-building by designing and coordinating a network of European skills centres for big data analytics technologies and business development beside others

2.1.4 Impact on societal challenges

“Europe 2020 is the EU's growth strategy for the coming decade. In a changing world, we want the EU to become a smart, sustainable and inclusive economy.” notes J. M. Barroso⁸⁶. These goals are reflected by the Horizon 2020 Framework Programme by focusing on seven distinct Societal Challenges⁸⁷. Although QROWD could not contribute to all objectives, our business cases, which are used to validate our concepts, are explicitly aligned with the following challenges.

Secure, clean and efficient energy: the energy challenge is designed to support the transition to a reliable, sustainable and competitive energy system.

The REEEP's business case of the open data and knowledge services (cf. Section 1.3) is strongly related to this challenge as it is the objective to provide sophisticated data analysis and knowledge management tools to easily navigate renewable energy and energy efficiency concepts, terms, and related or similar documents. Due to the multilingual nature (English, French, Spanish, Portuguese and German) of the thesaurus and the corresponding definitions it is an outstanding foundation for robust decision making processes in governmental and non-governmental institutions as well as corporate companies in Europe and all over the world as for instance: World Bank, UN Data, UN-REDD, UNEP, GGKP, NREL, Eldis, The German Government & GTZ, weADAPT, Energypedia, Energy and Environment Finland or Climate Tech Wiki, FAO Stats, REN21 or CDKN in UK.

A critical threat for REEEP's success is continuous improvement – regarding the quantity and the quality of the information – of the content pool to guide knowledge brokers through reliable and timely information. Thus, the advanced methods and tools developed or extended in QROWD will have a huge impact on the business case and, consequently, on the mentioned societal challenge. They allow for continuous harvesting, scraping, and publication of up-to-date clean energy businesses and stakeholders within European community. Moreover, the enhanced tool chain enables the trends analysis of crowd-sourced data so policy makers can obtain on demand reports of real time data for their decision making. In the end, policy makers will have a first class one stop shop for statistics and trends on clean, efficient energy technologies and country profiles for comparative analysis to underpin current policy frameworks.

Climate action, environment, resource efficiency and raw materials: activities in this challenge will help increase European competitiveness, raw materials security and improve wellbeing. At the same time they will assure environmental integrity, resilience and sustainability with the aim of keeping average global warming below 2° C and enabling ecosystems and society to adapt to climate change and other environmental changes.

Similar to the explanation above, REEEP's business case also contributes to this societal challenge since the offered, multilingual thesauri enable the knowledge workers to navigate climate change concepts, terms and related

⁸⁶ http://ec.europa.eu/europe2020/index_en.htm

⁸⁷ <http://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges>

or similar documents in the field of climate compatible development. Furthermore, the Reegle content pool on clean energy and climate change development is constantly growing and comprises comprehensive and sustained global environmental observations and information. Unfortunately, it becomes more and more challenging to keep the data and information quality due to the increasing size of the knowledge base. Hence, the results of QROWD will significantly improve the quality assurance processes. This will ensure and advance REEEP's market position, also in the fields of Open Data Consultancy Services. The latter address the vast growing number of information portals for climate change adaptation and mitigation (e. g., <http://www.pacificclimatechange.net/>, <http://www.un.org/climatechange/>, <http://www.unep.org/climatechange/>), which work with unmanageable webs of documents and data behind them. All in all, through REEEP's participation in QROWD key resources in the climate change mitigation and adaptation fields will be more productive and effective having the knowledge required at their fingertips, not replicating work of others or spending too much time finding relevant documents and data.

Smart, green and integrated transport: This challenge aims to boost the competitiveness of the European transport industries & achieve a European transport system that is resource-efficient, climate-and-environmentally-friendly, safe & seamless for the benefit of all citizens, the economy and society.

A particular area of competence of BROX is supporting automotive supply chains. BROX has worked on a number of projects in this area with Volkswagen, Daimler and other customers. In particular, BROX is member of ITA - Automotive Service Partner, the German Association of the Automotive IT industry and in this role involved in standardizing crucial IT building blocks and defining best practices. Quality assessment as provided by QROWD is relevant for supply chains in two respects:

- As supply chains are large and contain data coming from many suppliers from different countries, assessing and maintaining the quality of multilingual data, such as customer reference data, is essential.
- Optimizing supply chains usually requires the integration of external background knowledge, such as route maps. When using external data sets as background knowledge, it is crucial to assess the quality of these, to complement and contrast it with additional data being available to the enterprise (e.g., enterprise taxonomies, domain databases etc.).

Quality assessment helps to make supply chains more efficient and flexible reducing transport and storage costs and by that also reduce CO² emissions. It is estimated that better decision-making by travelers, but also transport operators and authorities, as well as businesses seeking logistics support, can help reduce CO₂ emissions levels with up to 380 million tones worldwide.⁸⁸ BROX leading position as a technology provider in the automotive sector will have a standing in this space.

2.2 Measures to maximize impact

2.2.1 Dissemination and exploitation of results

The centre of the innovation action on hand is to improve the competitive advantages of all partners in the QROWD project. The whole project design is built around the business cases, which are the basis for research and development activities carried out by the project (see also Figure 2.2.1). So exploitation is not only an additional effort to the technical development, but it is the basic principle we are paying attention to on every stage of the project. A dedicated Exploitation Coordinator (see also 3.2.1.6 Exploitation Coordination (ECO)) ensures this, not only in the already planned tasks and deliverables, but also utilizing emerging exploitation opportunities where they arise.

Dissemination and exploitation in QROWD includes the project itself as well as mainly the major results and outcomes during the project but also beyond.

By taking up the results of the project - in this case the data cleansing and repair services developed in WP1 - into commercial products and services, all European citizens can be reached while ensuring profitability through economies of scale. Additionally, volunteers will profit from free tools to facilitate data curation and improve the quality of the Linked Open Data Cloud. Data quality is one of major bottlenecks for the effective reuse of open and linked data sets. The availability of mature services in this area opens a diversity of exploitation opportunities, from Data-as-a-service (DaaS) and marketplace providers, data catalogs and repositories, as well suppliers of sector-specific solutions using linked data technologies (including dynamic semantic publishing and data integration).

⁸⁸ McKinsey. Big data: The next frontier for innovation, competition, and productivity.2012

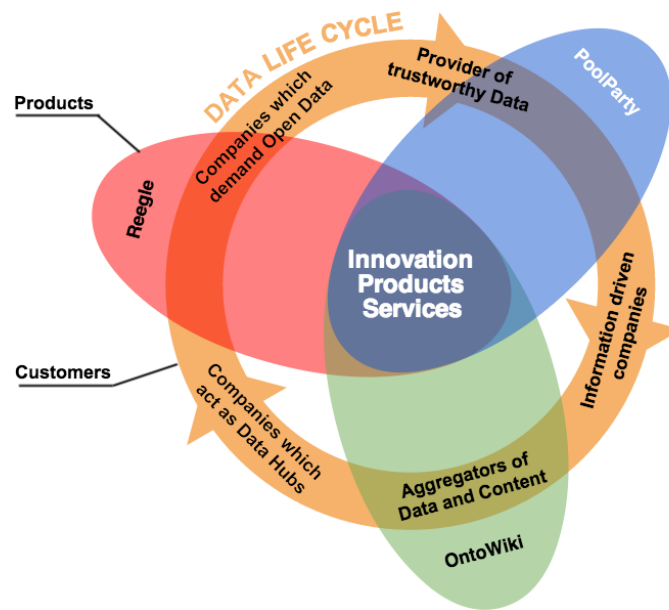


Figure 2.2.1 Overview of exploitation approach in QROWD

An integral part of the exploitation approach is the implementation of the two business solutions (WP3/4), which will serve as the validation point throughout the project. Wherever possible, research results will be exploited for the internal development and support of new products and services. These products and services will lead to a competitive advantage of the participating organizations and will substantially contribute to the benefit of the targeted users. In order for the exploitation to be effective, an integrated approach will be necessary, combining experience and expertise from the development department and the product ownership & management, as well as the involvement of a customer base / the market brought in by the industrial partners in the two designated business cases as well as the public service deployments.

2.2.1.1 (Research) data management approach and standardization

QROWD will gather and make use of data mainly in the areas of the two business scenarios, means in the fields of clean energy and climate change development as well as in eCommerce and finances using various linked and open data sets, many of which are released and maintained by public bodies.

Furthermore we will use the novel QROWD cleansing and curation mechanisms to be tested on free available LOD sources. Such data sources are e.g., DBpedia, LinkedGeoData, Freebase, Eurostat, furthermore GEMET and several data sets in the relevant domains of publicdata.eu.

The project will also make use of extended existing data as for instance the Reegle glossary⁸⁹ (a multilingual SKOS thesaurus on clean energy) as well as create new data as knowledge models and ontologies like SKOS Thesauri that partly will be published as linked open data with an open license (as e.g., a Creative Commons Attribution) to be used and reused by all interested parties for free. Regarding the used standards all data sets are available or will be made available as Linked (Open) Data following the respective recommendations / standards of W3C.⁹⁰

Data Management Plan (DMP)

A DMP is a document outlining how research data will be handled during the QROWD project, including the following issues:

- What types of data will the project generate/collect?
- What standards will be used?
- How will this data be exploited and shared/made accessible for verification and re- use? If data cannot be made available, explain why.
- How will this data be curated and preserved?

⁸⁹ <http://www.Reegle.info/glossary/>

⁹⁰ <http://www.w3.org/standards/semanticweb/data>

The described policy should reflect the current state of consortium agreements regarding data management and be consistent with those referring to exploitation and protection of results. The first version of the DMP is expected to be delivered within the first 6 months of the project. This DMP deliverable, which is part of the Work Package on Project management (WP7), should be in compliance with the template provided by the Commission (see Annex to “Guidelines on Data Management in Horizon 2020, Version 1.0, 11 December 2013, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)

The data created through crowdsourcing will be published under a license compatible to the one originally set by the provider, in order to facilitate reuse. In WP1 we will explicitly look into this issue and make a proposal for a purposeful model for the newly created or curated data. Licenses of harvested data sets will not be changed – where ever needed the consortium will manage a license clearing process when for instance integrating two data sets with different licenses in place. We will participate in the Open Research Data Pilot initiated in Horizon 2020 and develop a Data Management Plan describing our involvement.

We will analyze user data that allows us to learn profiles of crowd contributors. This does not include login information, but focuses anonymized interaction data. For crowdsourcing campaigns run within the verticals, full access to personal data will only be granted to system administrators. QROWD has no intention to give away, sell, or exploit this data in any way. Learning about crowd performance relies solely on interaction data and not on personal details.

Standardization

Regarding standardization activities the QROWD consortium is well involved in several standardization activities at W3C, the ISA programme as well as national standardization bodies and activities. The representatives of the QROWD consortium will bring the project results – mainly mechanisms of data quality control and assurance – into this standardization bodies as well as bring the latest developments from the standardization bodies into the project to be evaluated and – where applicable – used.

QROWD partners are already very active in standardization activities:

- Martin Kaltenböck (SWC) is W3C invited expert, invited experts of the Cooperation OGD Austria (national body) as well as working in ISA programme working groups.
- Daniel Hladky (Ontos): W3C Business Development Representative in Switzerland and Russia. Ontos is also member of the W3C community group “GeoSemWeb”.
- Martin Voigt (Ontos): is Co-Chair of Semantic Web Interfaces Community Group
- Florian Bauer (REEEP): is co-founder of the Climate Knowledge Brokers Group (CKB) & partners in Linked Open Data and Sustainable Transport project development with W3C

The responsibility for future data curation of the mentioned data sets (also beyond the project duration) lies on the side of the respective business scenario owner / operator that makes continuous use of the data means REEEP and UNISTER.

Knowledge and IPR management

The consortium partners agreed to publish an important share of the software created or extended in QROWD under open source licenses. The crowdsourcing services developed in WP1 will be openly available under GPL3 or Apache or similar licenses.

Regarding the products of the three ICT vendors, OntoWiki knowledge engineering and application development platform uses a GPL2 license; PoolParty Semantic Suite and OntosLDIW can be used under commercial license using a subscription-based model, while the former offers free or cost-reduced licenses for academic or educational purposes.

In order to ensure a smooth execution of the project, the project partners agree to grant royalty-free access to Background and Foreground IP for the execution of the project. Such and similar issues will be addressed in detail within the Consortium Agreement between all project partners, which will define the legal framework required to ensure that the project is maximally beneficial for the individual partners and the consortium as a whole and the goals of the project. In WP 6 Impact Creation an IPR strategy and plan will be developed together with all partners as an integral part of the exploitation plan.

The focus of the project is on technology transfer. As such the publications produced by the consortium will have an applied character and describe case studies and experience with the technology. By default we will follow a

green open-access model and publish and self-archive all our publications using the ePrints service of the University of Southampton.⁹¹

2.2.1.2 *Collaboration with other projects*

Collaboration with other (EU funded) projects is two-fold: on the one hand QROWD will make use of the outcome and results of other projects as for instance of UnifiedViews⁹² that was developed in the course of the LOD2 project⁹³ by SWC and University of Economics in Prague, Czech Republic. On the other hand QROWD will investigate to disseminate the projects results to be used by other projects.

2.2.1.3 *Dissemination and exploitation activities during the project*

Continuous dissemination activities along the whole project duration along the dissemination plan that will be developed in quarter 1 of the project in WP6 Impact Creation and will be continuously updated in bi-annual periods during the project duration. Concrete activities are as follows, but not limited to

- Project Web site
- Print material on QROWD (print on demand) as leaflet and sticker, a roll-up
- Make use of partners communication channels
- Make extensive use of existing mailing lists (e.g., W3C, OKFN, STI)
- Make extensive use of social media as Twitter and LinkedIn to promote project results
- Manage PR (public relations) activities as partner press releases on the new products and services and event organization / participation presenting new products and services – always mentioning QROWD as the underlying project.
- Event participation on behalf of QROWD (presentations, exhibition stands, workshop participation), but mainly regarding the data quality components, the 3 extended data management products and the two business solutions developed in QROWD.
- Academic publications and presentations of the data quality and curation approaches of QROWD by the two academic project partners.
- Continuous analysis of exploitation opportunities, adjusting the project when necessary in order to ensure the best possible outcome as well as exploitation activities along the exploitation plan that will be developed together of all partners in WP6 in quarter 2 of the project and a report on dissemination activities will be delivered in month 30. Each individual SME partner will have individual extensions of the exploitation plan related to their market and product activities.
- Investigation into the possible economic benefits and impact of the expected results. Continuous evaluation of the advancement of the research results against the user requirements/needs throughout the project with the help of the user partners and adjustment of the project when necessary (see also the market validation approach in Section 1.3).

Additionally we will showcase the QROWD data quality services for a core of the Linked Open Data Cloud (such as DBpedia or data sets of publicdata.eu) and openly deploy them to allow data publishers and consumers to test and use them in order to solve their own data quality problems.

2.2.1.4 *QROWD innovation to market approach: customer segmentation and economic effects*

To develop a view on the customers segments QROWD aims to address, we have to consider, that the business applications of data management are based on companies utilizing new and improved insights. According to a recent study⁹⁴, the resulting economic gains can be put into three broad categories:

- Resource efficiency improvements through reducing the information concerning resource waste in production, distribution and marketing activities,
- Product and process improvements through innovation based on R&D activities, day-to-day process monitoring and consumer feedback,
- Management improvements through evidence-based, data-driven decision making.

⁹¹ <http://eprints.soton.ac.uk/>

⁹² <https://github.com/UnifiedViews>

⁹³ <http://www.lod2.eu/>

⁹⁴ Report “Big and open data in Europe: A growth engine or a missed opportunity?”, <http://www.bigopendata.eu/>

The channels through which efficient data management using open data sources affects the economy are closely linked to the aforementioned categories. There are three ways in which data can turn into valuable input for a company:

- Data-to-Information is the key source of data value for business models where information is the core source of value. The data is “mined” for valuable information and the value is created the moment the search succeeds.
- Data-to-Product/Process effects occur when insights from data analysis need to be implemented in the physical world to bring value to an enterprise.
- Data-to-Management systematically brings database information into a company’s decision-making processes.

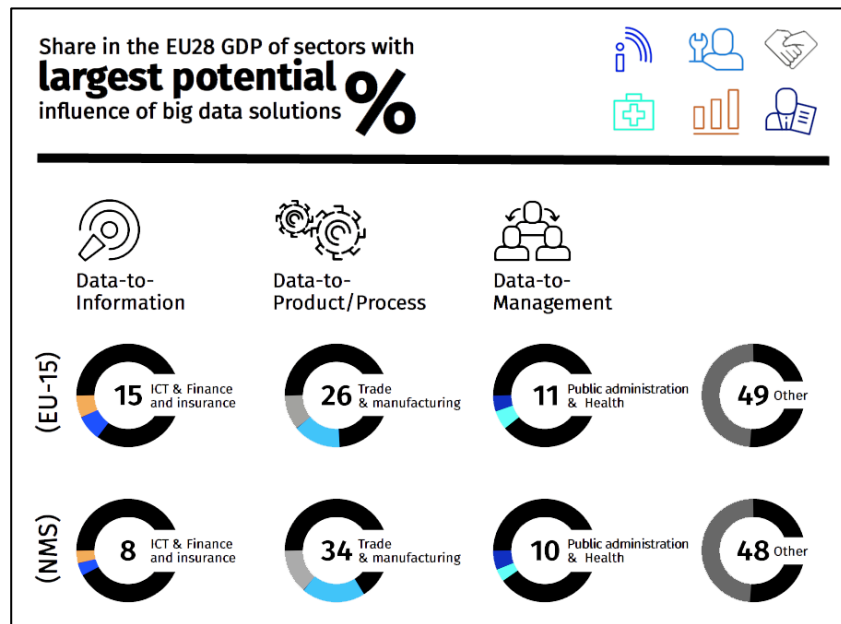












Figure 2.2.2 Exploitation of data assets

Figure 2.2.2 visualizes how these three channels affect different sectors of the economy disproportionately. Data-to-Information is especially important for the financial and insurance sector as its business models depend on gaining advantage from the information available on the financial markets and potential consumers respectively. The eCommerce and other areas of ICT-based economy are inherently dependent on maximizing the value of information available to them. Data-to-Product/Process will bring innovation and efficiency gains to manufacturing through better insights from various sources, including rapidly advancing sensors. Brick-and-mortar wholesale and retail trade can benefit from supply chain and inventory optimization based on big data analysis. Data-to-Management is broadly applicable across various sectors, but it can have the biggest impact on public administration and healthcare which are especially prone to inefficiencies due to lack of readily accessible information on performance and optimization opportunities.

Taking this data management macroeconomic picture to the consortium’s reality, following customer segments are targeted with QROWD in either a direct vendor-customer relationship or via a liaison over a third party.

	Customer segments relevant to QROWD	Targeted by				
		SWC	UNISTER	BROX	REEEP	ONTOS
 Data-to-Information	Information driven companies, which act as a hub for domain- or theme specific taxonomies, vocabularies, glossaries and thesauri.	✗				
 Data-to-Management	Corporate companies with a disperse structure, several branches (multinational, multi domain global scale corporates)	✗		✗		✗
 Data-to-Information	Companies which act as agency for trustworthy data for third parties (financial, health, pharm, environment)	✗		✗		✗
 Data-to-Information	Governmental and non-governmental institutions acting as reference point for information and structures of their domain (branches associations, statistic bodies)	✗			✗	✗
 Data-to-Product/Process	Data-driven companies with need for high quality data or recent data, e.g., online booking agencies		✗	✗		
 Data-to-Product/Process	Product-driven companies that need a distribution channel for their product or an enrichment for their data, e.g., hotel operators		✗			
 Data-to-Information	Information-driven companies that need access to trustworthy information about a specific topic, e.g., market research, consulting companies		✗		✗	
 Data-to-Information	Public and private international organizations working in development field such as Governments and non-Government organizations, Multi-lateral organizations (existing, current) receiving free open data to enrich own data and own Web presence	✗			✗	✗
 Data-to-Product/Process	International organizations based in Europe who require a Tagging API to assist in categorizing documents, connecting to related and suggested documents in the Clean Energy, Renewable Energy and Energy Efficiency fields in order to improve productivity, increase interoperability and collaboration as well as categorize currently unmanageable libraries of information management documents and data				✗	
 Data-to-Information	Customers, which have to create on-demand reports of customized data and structured statistical data from the Reegle Content Pool / Tagging API ⁹⁵				✗	

2.2.1.5 QROWD dissemination and exploitation activities after the project

As the project is based on a strong business interest of all partners the exploitation of results will not stop at the end of the funded projects. The business development of the industry partners will take over these efforts for a partner-specific exploitation to ensure a successful market entry. In the course of the project the market validation will be carried out (see: section 1.1.3.) but the market entry will be managed after the project duration.

The following concrete dissemination and exploitation activities are planned by the QROWD industry partners to be managed after the project duration.

⁹⁵ <http://api.Reegle.info/>

REEEP

- Relevant tools and resources developed by QROWD fostering the business scenario of REEEP will be exploited through existing International Development Network and at presentations / workshops in European based conferences where tools will be demonstrated along with EU Commission branding as key funder
- Implemented Quality Assurance mechanisms for the Reegle Tagging API and Content Pool which will be reviewed on a scheduled basis and continuously improved by operational resources and future funding opportunities: results will be exploited in reports to existing and future customers to articulate value added as a result of this project.
- Implement Feedback loops from users and within the system for the Reegle Tagging API will be implemented to ensure continuous improvement of the tool which can also be exploited as a continuous value to end users, existing and future customers and will be exploited through marketing channels
- Implement Out of the box CMS integrations for Reegle API Tagging and pushing content into the Reegle Content Pool for API adopter organizations, exploiting the ease of use the tool offers in parallel with innovative and effective solutions for their business and development needs
- Enhance multilingual Reegle Tagging API Services for quality check and appropriate tagging in multiple languages which will go beyond the project and be included as value add in marketing material distributed in conferences, individual consultations and proposals for future development
- Continuous funding for further development and expansion of the Reegle Thesaurus, crossing further sectors and borders
- Continued multilingual publication of new thematic areas within the Reegle Thesaurus
- Implementation of Thesaurus Structure usage for enhanced tagging modelled off hierarchy and related terms, not just frequency of terms
- Marketing and consultation regarding enhanced tools for continuous adoption of the Reegle Tagging API by additional organizations in the field
- Integration and analytics in a continuous data value chain process consistently collecting, cleaning and publishing new linked open data sources to the Web to assist in organization of internationally required data in order to solve global problems related to renewable energy, energy efficiency, and climate change
- Generate usage reports, including which organisations are pushing content to the Content Pool, which organizations are using the Reegle API Tagging Tool, frequency and volume of usage to be exploited in marketing, consultation services and further development of the tools beyond the project

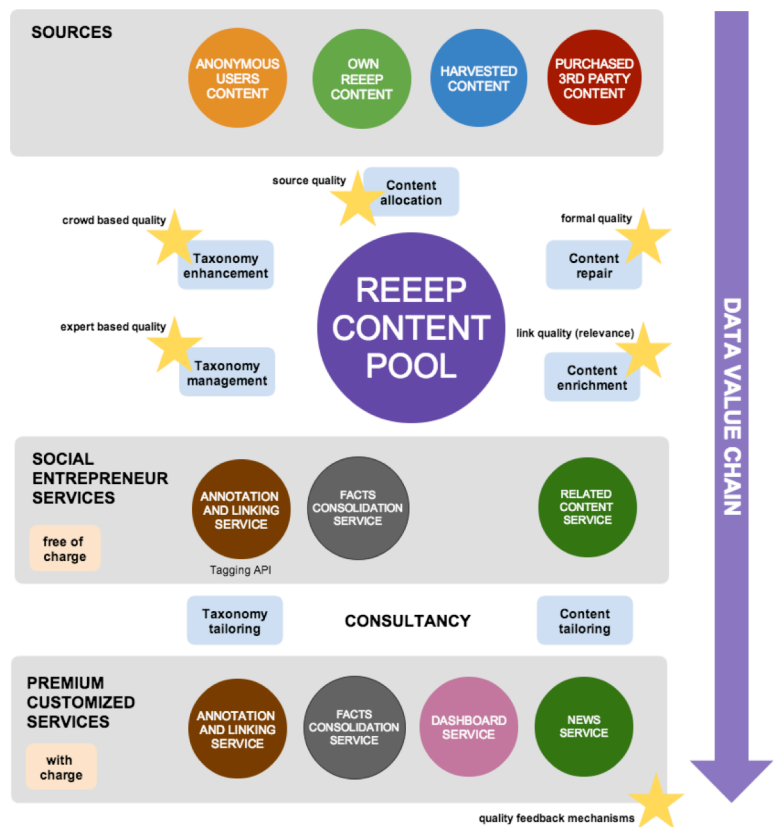


Figure 2.2.3 Relevant tools and resources developed by QROWD fostering the business scenario of REEEP

UNISTER

- Relevant tools and resources developed within the QROWD project fostering the business cases of Unister will be exploited through existing distribution channels within the business domains where Unister is already active. Several services for data querying and exchange are available at:

- The online and offline travel industry (Unister is a major player at the segment of booking vacation packages as well as flights within Germany),
- The retail shopping and comparison via Web applications, and news publishers.
- Implementing a quality feedback loop within the currently existing services for providing high quality feedback in very short time
- Provide multilingual data sets enriched with specific business domain-specific information
- Quality enhanced publication of sentiment data sets (aggregated user opinions) for products
- Apply standard vocabularies to imported public and private data sets to increase the reusability of the data sets
- Establish new data-driven applications in particular within the local driven field of mobile applications that is highly depending on a vast number of high quality data items.
- Driven by the exchange of thoughts and ideas with business developers and startups extend the number of services (Software as a Service) for providing specific parts of open data that are consolidated from public and private data and providing information about specific entity types like museum, home entertainment or administrative institutions.

SWC – Pool Party

Several standards exist for expressing controlled vocabularies but with the wide adoption of the linked data concept, the Web-based SKOS data model has become the choice of many contributors who want to integrate their controlled vocabularies into an interconnected Web of Data.

User need driven exploitation

In addition to various automatic quality checks made either in the background or as a batch job, which are already implemented into PoolParty (see Figure 2.2.4), customers will get additional plugins (marked with yellow stars) which give a technical answers, to customers demands in the area of enhanced data quality mechanisms.

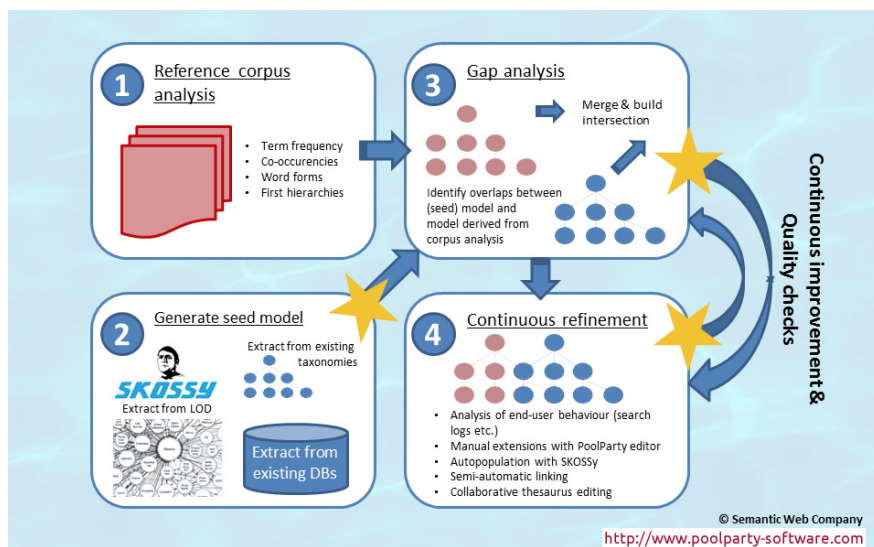


Figure 2.2.4 Existing quality checks in PoolParty

While the general location of plug-in design is already defined, exploitation activities will be driving a process of fine-tuning and customer request achievement. The envisaged plugins developed by QROWD serving the following user’s needs:

- Quality enhancement for text corpus analysis
- Quality inter-linkage to linked open data
- Quality proofed entity extraction and tagging

Exploitation activities will be based on the already proofed system of possible measures carried out by SWC in regular customer requirements elicitation processes:

- Showcase (SC): Following the customer and use case analytics (see above) we will build prototypic Web services for the general (free) use, which demonstrates the power of QROWD based PlugIns for Poolparty.

- MockUps (MU): On basis of expressed customers needs, we build user interfaces that show the end user what the software will look like without having to build the software or the underlying functionality. MockUps can range from very simple hand drawn screen layouts, through realistic bitmaps, to semi functional user interfaces.
- Proof of concept (PoC): Together with potential customers a rough prototype of a new idea (customer request) is constructed. as a "proof of concept". This includes a) rough requirement specification, b) an agile description of s/w design, a prototypic user interface and an ad-hoc test phase. At the end of the PoC customer may order a serious integration or a SotA software development.

ONTOS – Linked Data Information Workbench (LDIW)

The linked data paradigm has been well established in the academic world and since the introduction of the linked data lifecycle by the LOD2 project the process is very well defined. Ontos view is to create an impact in the enterprise world and enhance the research results from LOD2 and GeoKnow through the QROWD project by enhancing identified technology areas. For this purpose we will use the existing GeoKnow workbench and the Linked Data Stack in order to address the following shortcoming:

- Improve the data quality with enhanced cleansing and repair functionality
- Through crowdsourcing methods and services improve linking of data sets
- Provide gamification tools that inspire user to provide a better contribution for quality assessment, repair and linking

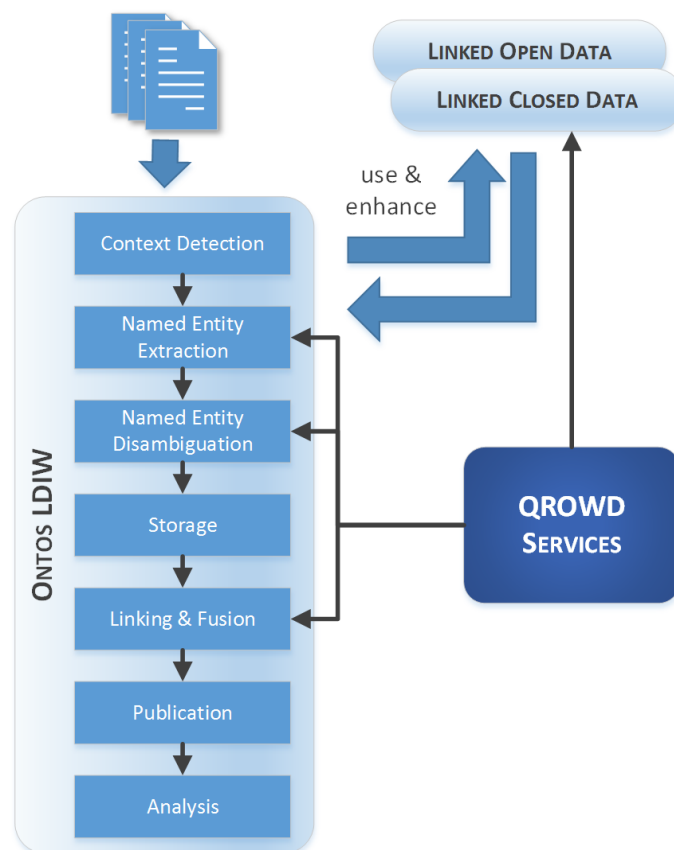


Figure 2.2.5 Exploitation strategy of Ontos in QROWD

The above-mentioned improvements will be fully integrated into the OntosLDIW (WP3). The QROWD services will improve the data value chain of the OntosLDIW (see Figure 2.2.5). Ontos will validate this functionality with two major use cases within their customer base. 1. Improved News aggregation by enhancing the named entity recognition and interlinking; 2. Data integration of external (open) and internal data sets in order to mashup the CRM System (Customer Relationship Management). Following Key Performance Indicators (KPI) will be used to measure the result:

- KPI – Link Quality > 85%
Establish a gold standard to measure the improvement through the QROWD project.

- KPI – Crowd contribution vs. internal 60: 40
60% or more external contribution to improve quality, repair and linking. Measure the number of people using the crowd service and gamification.
- KPI – Revenue growth by 30%. Exploiting the result to the market should lead to an increase of the current Ontos revenue of at least 30%.

The Ontos Business Model (see also Ontos business canvas in section 1.3) foresees the impact in three major areas:

- Information Integration within Enterprises from Medium to Large size organizations. A growth rate of at least 30% in this area is expected.
- Establish a partner eco system with system integrators that will address the market. The revenue from this reseller model shall contribute at least 20% of the overall Ontos revenue.
- Deploy the solution to ministries that like to do information integration and open data publishing. Switzerland and Russia are the prime markets for this and the revenue in this area should represent at least 25% of the overall Ontos revenue.

The go to market relies heavily on demonstrating the solution through Web demos, at conferences and exhibitions. A key resource to achieve this is a database with qualified contacts that can be converted to leads and customers.

BROX – OntoWiki / eccenca Linked Data Suite (eLDS)

BROX offers the commercial eccenca Linked Data Suite (eLDS), which is based on open source components that primarily have been developed in the LOD2 EU FP7 project.

Improvements to OntoWiki

eLDS includes OntoWiki extended with commercial add-ons as its Data Wiki, which means the improvements to OntoWiki done in QROWD will flow into eLDS. As an essential asset of its core business, the open source version of OntoWiki will be continued to be maintained by BROX in close collaboration with its primary maintained INFAL. As part of eLDS, OntoWiki is already installed at major customers of BROX, such as Volkswagen and Daimler, for managing internal knowledge bases.

Quality assessment

Assessing and maintaining data quality is a major concern for many of BROX customers. Currently, eLDS does not include a quality assessment component as no existing open source software that fits all customer needs could be identified in the linked data space. The improvements for quality assessment that are made available by QROWD will close that gap. A quality assessment component will increase the value of eLDS considerably.

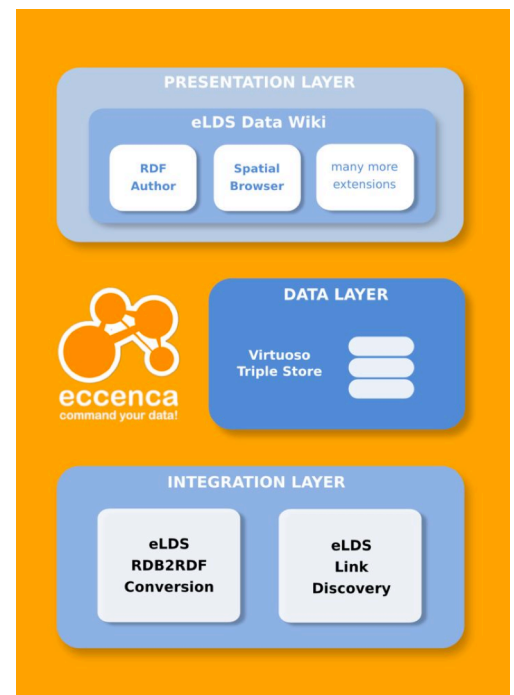


Figure 2.2.6 BROX technology taken up in QROWD

2.2.2 Communication activities

The project will take a pragmatic approach to results' communication, considering dissemination instrumental to exploitation and as a vehicle to facilitate and support it. As such the major activities will be: (i) public promotion material, online presence and use of social media / networks; (ii) industry relationships and showcasing activities; (iii) publications and participation at scientific events; (iv) training & tutorials on the use of the tools and the business solutions. These communication activities will be explained in more detail in the following.

2.2.2.1 Innovation, knowledge transfer and stakeholder integration

In general, we estimate to maximize impact of the QROWD project in Europe since the project partners have excellent background knowledge and established networks according these fields of data management and the data economy and have access to broad vertical markets, e. g., eCommerce (for instance UNISTERS travel portals) or

clean energy (for instance REEEPs clean energy portal Reegle.info). But also through the respective networks of all partners we identified the following target groups and stakeholders from corporate companies, governmental and non-governmental institutions:

- Data, information, and taxonomy provider and publisher
- Experts collecting and structuring domain knowledge
- Customers integrating the services of project partners
- Researchers in the fields of QROWD
- Policy makers

By intense communication with these stakeholders they will gain manifold profits from projects innovations, in particular:

- Easy integration of (semi-)automatic, platform-independent services for quality assurance into existing or future data value chains
- Structured, interoperable, interlinked, trustful, and multilingual knowledge bases
- Cost effective language translations that are reliable and break down linguistic barriers, border barriers, and sectorial barriers
- High-end services and high-quality data for competitive costs
- State-of-the-art methods and mechanisms for (Linked Open) data quality assurance deployed in a broad range of tools validated by three common, large-scale use cases with real customers

2.2.2.2 *Channels and activities*

In QROWD, we will make use of numerous channels to reach our target groups and transfer the knowledge gained during the project runtime. Therefore, all partners can build up on distinguished skills earned in European and national research projects and remarkable experience in the dissemination and exploitation of project results. Exemplary channels are:

- Own up-to-date products, services, and Web sites based on the high-quality, multilingual data
- Broad, existing, multi-domain and -national network of integration and sales partners
- Efficient consulting and support services in the area of (Linked) data quality assurance
- Descriptive talks and vivid demos at conference and exhibitions in the fields of Big Data, data and information management, data quality, and linked open data but also of vertical markets
- Own workshops and articles in high-ranked, maybe open access journals

Public promotion material, online presence and use of social media

We will establish a comprehensive online presence, including a project home page, a news section (including synchronized RSS feeds) and continuous updated information about the results of the QROWD project. The communication channels beside this will be the existing channels of all project partners, as well as research and technology based Web sites & social networks & media (e.g., SourceForge.net, LinkedIn, SlideShare, VideoLectures.net, Twitter), furthermore mailing lists (e.g., linking open data, semanticweb@w3c.org) and the blogosphere. Additionally, we will have small activities in state-of-the-art PR and marketing tools such as project flyers and brochures to be disseminated to events such as conferences, workshops, tutorials and industry events, and promote the crowdsourcing projects on platforms and blogs such as crowdsourcing.org and clickworker.com.

Industry relationships and showcasing activities

Each commercial partner plans to run at least one workshop per year where they will invite existing and potential customers in order to advertise the tools and showcase the newly developed functionality (this will also be used for market validation of the new products and services). In addition, they will have an active presence at various technology conferences such as European Data Forum, OKFest / OKCon, SemTech (EU/US), STRATA, and technical fairs such as Infa. The research partners SOTON and InfAI participate in a wide range of technology transfer and academic events such as ISWC, ESWC, SEMANTiCS or the Leipziger Semantic Web Days, and will participate in various hackathons (in collaboration with organizations such as the Open Knowledge Foundation and the Open Data Institute ODI)⁹⁶.

Given the importance of data cleansing and repair for the uptake of linked data, the results of our project will be at the core of the agenda of these events, which are attended by hundreds of institutions in Europe representing all

⁹⁶ <http://www.theodi.org/>

stakeholders of the data value chain. In addition to these partner-specific activities, QROWD will develop a tutorial and present it at conferences such as SemTech, SEMANTICS and LinkedData & SemanticWeb and OpenData Meet-Ups at the local and regional levels on the potential of crowdsourcing as a complementary means to deal with limitations of automatic linked data management techniques. We believe such a resource will be an important instrument to operationalize the usage of crowdsourcing as a tool for various types of tasks related to linked data, consolidating the experiences made in this, but also in other projects, on this topic to facilitate effective knowledge transfer and a positive evolution of this promising field. In relation to WP1 the consortium will run open challenges to seek the public opinion on the most stringent data quality problems in the community (see Section 1.3 and WP1 T1.3) and to encourage participation in open challenges and campaigns (similar to the one mentioned in earlier sections run by InfAI in connection to DBpedia). We will run three such challenges for three data sets covered by T1.3 starting from the second quarter of the project.

Publications and participation at scientific events

Another communication channel targets public scientific events, where the work of QROWD will be distributed in form of publications. They will address three main scientific communities, namely the ones around Semantic Web/linked data, social computing/crowdsourcing and Human Computer Interaction (HCI). In particular, we will publish the lessons learned from designing and applying the crowdsourcing services developed in WP1, as well as the methods for workflow and task management, which provide new insights in crowdsourcing research and the ways the overall ideas could be employed in knowledge-intensive scenarios. The tutorial mentioned earlier will also be presented at two scientific events (ISWC/ESWC and K-CAP /EKAW).

Trainings and tutorials on the use of the tools

To enable the target groups to learn about the tools developed in QROWD and to use them most efficiently a strong focus of the planned activities in communication lies in creating easy-to-understand material how to use the tools & services as screencasts, a webinar series, online tutorials and manuals as well as FAQs (gathered from mailing lists and above). This material will be created along a set of user stories that will be developed in the course of the initial dissemination planning & activities - user stories can be seen as story boards of several user groups / target groups that will be using the results of QROWD (components, 3 enhanced ICT products and the 2 business solutions). Additionally, introductory talks and screencasts of the business cases, will be created and published on youtube.com and existing groups & forums will be used and where applicable QROWD-related groups and forums will be started on social networks such as LinkedIn, Twitter or Google+ etc.

3. Implementation

3.1 Work plan — Work packages, deliverables and milestones

3.1.1 Overall strategy of the work plan

The work plan is scheduled over a period of 30 months and reflects the concentrated effort needed to achieve the project objectives. The project is planned to pass through five interconnected phases as shown in Figure 3.1.1:

- **Phase I:** Process and technology development for crowdsourcing - deals with the design and the realization of crowdsourcing services aligned with automatic approaches for data quality assessment and repair along the linked data lifecycle (WP1) [M1-M18];
- **Phase II:** Design and development of a data quality service ecosystem – is concerned with the creation of a set of services that bring together state-of-the-art automatic methods for data curation with crowdsourcing capabilities as part of the data value chain (WP2) [M7-M24];
- **Phase III:** Tool extensions - WP1 and WP2 results will enhance existing tool suites that are part of the Linked Data Stack with advanced quality assurance and repair services combining human and computational intelligence (WP3) [M7-M24];
- **Phase IV:** Application, testing, and evaluation - we will build data value chains for the two business cases (clean energy and eCommerce) on top of the technology delivered in WPs 1, 2 and 3. We envision an iterative process in which this technology will be customized to match the technical and process-related requirements within real-world business cases. This will also involve lab evaluations in terms of usability, accuracy, costs, and scalability at M18 and M30 (WP4 and WP5) [M4-M18];
- **Phase V:** Market validation - refers to the roll out of the technology in real customer application contexts, monitoring the performance, documenting the benefits, extending existing business models and measuring the impact in two iterations at months M18 and M30 (WP4 and WP5) [M4-M30].

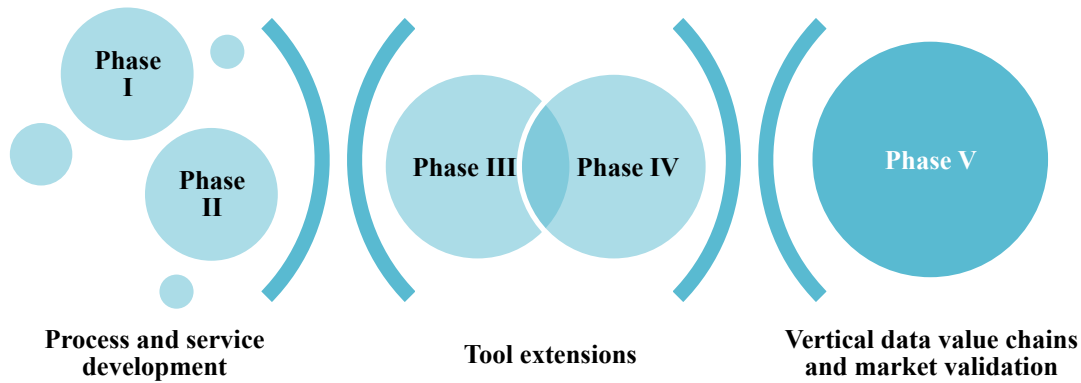
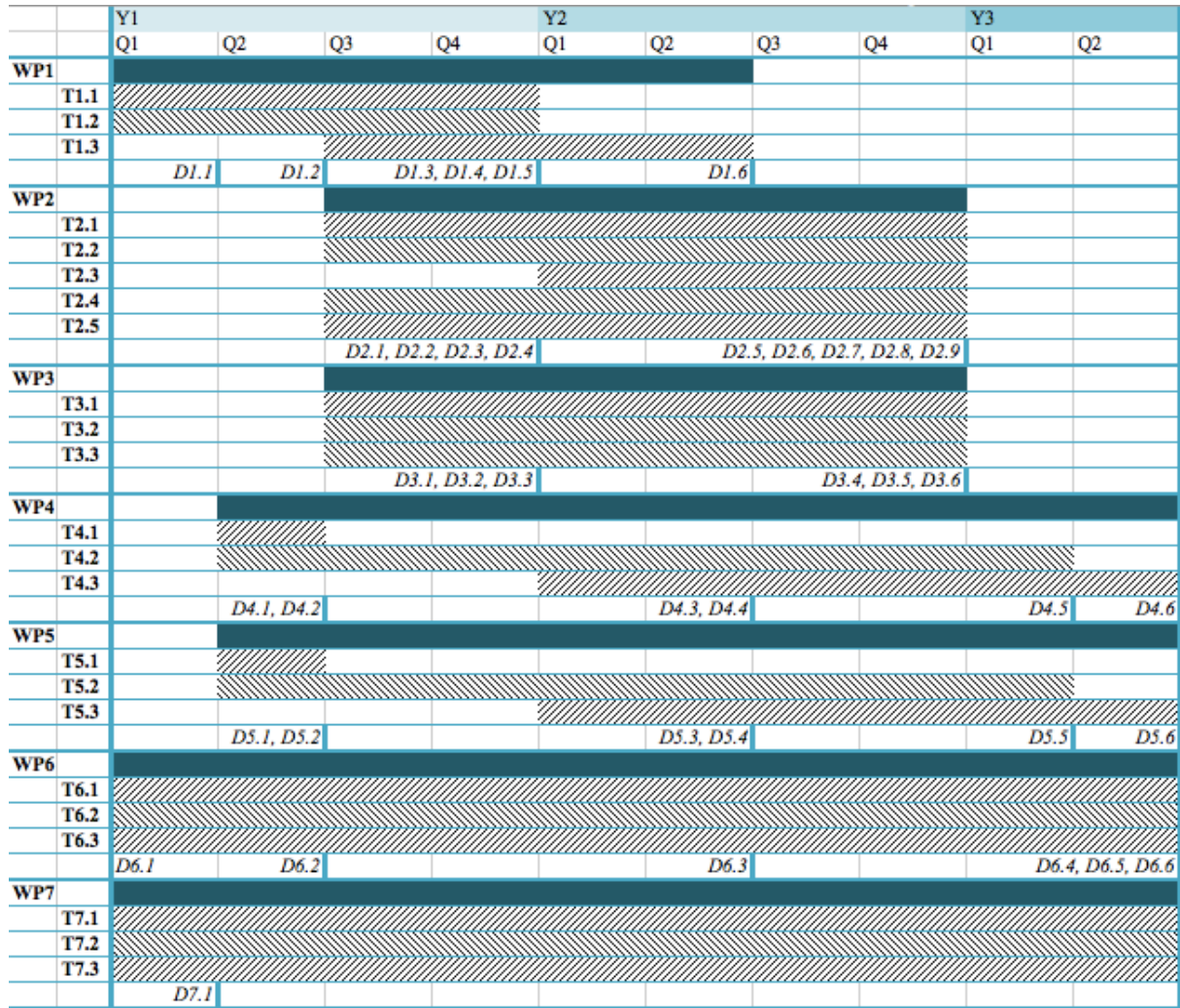


Figure 3.1.1 High-level phases of the work plan

QROWD is intended to be a highly focused project, producing high-impact contributions to linked data curation by combining crowdsourcing and automatic approaches. The project is structured into seven work packages. WP1 is concerned with research and development of crowdsourcing services for quality assessment and repair in the data value chain; WP2 with the creation of a service ecosystem, which brings these newly created crowdsourcing services together with automatic data curation methods creating a modular and extensible service ecosystem that spans over the entire data value chain; WP3 with the exploitation of these services in existing tool suites brought in to the project by the QROWD technology companies; and WPs 4 and 5 with the transfer of the developed data value chain technology into business verticals (clean energy and eCommerce). The technology will be further customized and deployed in productive environments. By evaluating the methods on real-world data sets, as an integral part of publishing and management tools that are used in various application scenarios QROWD has the potential to achieve a significant impact in the linked data technology landscape, in which one can observe a growing trend towards DaaS and other data-driven business models. The straightforward, yet effective nature of the innovation methodology we introduced, which relies on a mix of generic components and specialized functionality that meet the requirements of vertical domains, will create direct opportunities for exploitation in the two industrial sectors chosen for market validation (renewable energy and online travel purchases). These are described in Section 1.3. In addition, it will allow technology providers such as SWC, ONTOS, BROX and UNIST to add a diverse mix of curation services to their products, responding to a clear and increasingly pressing need to improve the quality of existing open data sets.

The work plan follows an iterative technology transfer approach, with first releases of each development piece released early and continuous evaluation based on feedback from direct customers and the broader community. WPs 1,2, and 3 each go through two iterations of their core deliverables – in WP1, as crowdsourcing capabilities are at the core of the data quality approach proposed by the project, they occur at M6 and M12. This allows WPs 2 and 3 to take up the results of WP1 and produce two releases of each service and tool suite, respectively, at M12 and M24. The second iteration will take into account the insights gained from field experiments, user studies, and community engagement (see also T1.3). Market validation activities start early and will be carried out for each of the two releases of the vertical data value chains. In Section 1.3 we provide further details on how the methodology followed in WPs 4 and 5 to introduce the new data-centric products into the market. The technology development and transfer work packages just discussed will be complemented by impact creation (WP6) and project management (WP7).

3.1.2 Timing of the different WPs and their components (Gantt chart)



3.1.3 Detailed work description

3.1.3.1 Work package descriptions

Work package description - WP1

Work package number	WP1		Start date or starting event				M1
Work package title	Crowdsourced data quality services						
Participant number	1	2	3	4	5	6	7
Participant short name	SWC	BROX	ONTOS	REEEP	UNIST	INFAI	SOTON
PM per participant	5	0	1	0	6	9	20

Objectives

In this work package we will first investigate those steps in the data value chain that are affected by data quality issues in order to establish quality assessment and repair as a general support activity accompanying every aspect of the data management lifecycle for Linked and open data. We will revisit existing frameworks and technology stacks in order to identify the best ways to enhance processes and their technical underpinning with quality assessment and repair capabilities. As far as this project is concerned, these capabilities cannot be performed but as a combination of crowdsourcing and automatic computation services. This work package covers primarily the crowdsourcing side of things (we say primarily because every meaningful crowdsourcing exercise that needs to scale to realistic data sets has to make some assumption about how the outcomes produced by the crowd feed into existing workflows). For the nine quality dimensions introduced in Section 1.1 we will thus propose specific types of crowdsourcing services, including microtask crowdsourcing (paid, with intrinsic and extrinsic motivation), gamification, volunteer campaigns, and open challenges (with modest prizes or reputation-centric) with different crowd audiences and platforms (professional environments such as CrowdFlower, but also less formal ones such as social networks). Building upon them we will look into two directions for further optimization, one related to how and to whom tasks are assigned, the other one minimizing delivery time through a different work load paradigm and explicit service level agreements with the participants. In WP2 they will be integrated into more complex data quality pipelines that will capitalize on state-of-the-art automatic quality assessment and repair techniques. These services will be integrated, possibly in an extended and customized form, in the data value chain technology environments covered by WP3, as well as in the two verticals (in WPs 4 and 5). Real-time crowdsourcing methods will be used in WP2 in T2.3. In addition, all services will be showcased for the general public in this work package; interested parties will be able to invoke and use them via carefully crafted Web APIs and test their openly available deployments. This will also give us a means to get comprehensive feedback from the broader community. As part of this effort we will define vocabularies and propose license models for publishing and exchanging the data created by the crowd, as well as basic procedures for the use of the public end-points.

Task 1.1 Design and implementation of crowdsourcing services (M1-M12; Lead: SOTON; Participants: UNIST, INFAI) This task is concerned with the specification of the types of tasks to be crowdsourced and the choice and parameterization of purposeful crowdsourcing solutions. For each type of task, we will support the following configuration parameters: inputs, outputs, as well as budgetary, time, and quality constraints. As an additional challenge we will examine the integration of crowdsourcing approaches with (semi-)automatic methods for data quality assessment and repair. This requires an analysis how these different approaches can be combined and impact each other, but also a principled approach to planning in the overall data value chain, due to the obvious differences between human-driven and classical (Turing) computation (i.e., performance, deterministic behavior). For microtask crowdsourcing and gamification we will propose several options for the breakdown of work into HITs, the design of human-readable interfaces, HITs grouping and anti-spam measures, qualification tests, and automatic evaluation methods (such as majority voting, statistical techniques etc.). For volunteering and open challenges we will do a succinct analysis of motivators and incentives (in WPs 4 and 5, for the public showcases in T1.3), and propose different mechanism-design-inspired methods that lead to the desired changes in the user behavior. To do so, we will use methodology devised in the INSEMTIVES project, to which some of the members of the QROWD team were involved,⁹⁷ as well as the experiments conducted by SOTON and INFAI on DBpedia.⁹⁸ We will implement standalone components for each form of crowdsourcing. For instance, for microtasks, we will have services invoking, e.g., CrowdFlower to add, modify, and translate labels, expand vocabularies and so on (see also Section 1.3 for additional examples of tasks). For games, we will

⁹⁷ <http://www.insemtives.org/>

⁹⁸ M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer and J. Lehmann. Crowdsourcing Linked Data Quality Assessment. The Semantic Web–ISWC 2013, 260-27, 2013.

develop a gamification API for managing profiles, scores, badges etc. We will use a combination of machine learning and user feedback and monitoring in order to identify those types of tasks which are likely to benefit more from curation by the crowd.

Task 1.2 Workflow, task, and time management (M1-M12; Lead: SOTON; Participants: UNIST) In this task we will look into more advanced methods to optimize the results of the basic services implemented in T1.1. In particular we will implement a method to deal with more complex workflows and coordination needs (following a customized version of TurkIT⁹⁹ and Turkomatic¹⁰⁰ for recursive tasks, built on top of clickworker and CrowdFlower), required to tackle open-answer tasks as in machine translation or label generation. Additionally, we will deal with the question of resource management, extending the profiling service of CrowdFlower and clickworker in order to improve the way tasks are assigned to workers. This is essential if microtask crowdsourcing, or other similar approaches, were to be applied in enterprise environments in order to identify appropriate expertise and request specific contributions. Task assignment will rely on a combination of existing profile information (for instance, as available within a company's employees repositories), and historical data about contributions to the given crowdsourcing project. Time aspects will be handled by implementing a so-called near-real-time crowdsourcing approach in which time constraints have a direct impact on the reward; alternatively, we will try out a model in which participants commit to certain service level agreements (e.g., availability and response times). The resulting methods will be integrated in a basic form in the services delivered in T1.1 and in showcases in T1.3, and in a domain-specific, extended version as part of the vertical data value chains realized in WPs 4 and 5.

Task 1.3 Public endpoints and service deployments (M7-M18; Lead: INFAl; Participants: SOTON) A first goal of this task is to design vocabularies and procedures, and propose license models for publishing the data created by the crowd. Vocabularies will cover content, process, and provenance-related topics, and reuse schemas such as VoiD. A second goal is to offer public services for the community to curate their data. The showcase will cover five distinct data sets, including DBpedia, LinkedGeoData, as well as three other data sets from PublicData.eu¹⁰¹ and the EU Data Cloud¹⁰². A final selection will be made at M4 following the recommendations of our customers in WPs 4 and 5, as well as of the general community (see Section 2.2). External parties will also be able to download and use the software to curate their own data.

Deliverables

- D1.1 Crowdsourcing design (M3; Lead: SOTON)
- D1.2 Crowdsourcing services v1 (M6; Lead: SOTON)
- D1.3 Methods for workflow, task, and time management (M12; Lead: SOTON)
- D1.4 Crowdsourcing services v2 (M12; Lead: INFAl)
- D1.5 Crowdsourcing vocabulary and licensing (M12; Lead: SOTON)
- D1.6 Public endpoints and deployment (M18; Lead: INFAl)

⁹⁹ M. Goldman, G. Little, L. Chilton and R. Miller. TurkKit: tools for iterative tasks on Mechanical Turk. In Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '09, 29–30, 2009.

¹⁰⁰ A. Kulkarni, M. Can and B. Hartmann. Turkomatic: automatic recursive task and workflow design for Mechanical Turk. In Proc. 2011 annual conference extended abstracts on human factors in computing systems, CHI EA '11, 2053–2058, 2011.

¹⁰¹ <http://publicdata.eu/>

¹⁰² <http://latc-project.eu/datasets>

Work package description – WP2

Work package number	WP2		Start date or starting event				M7
Work package title	Hybrid data quality services						
Participant number	1	2	3	4	5	6	7
Participant short name	SWC	BROX	ONTOS	REEEP	UNIST	INFAI	SOTON
PM per participant	8	12	10	0	12	22	9

Objectives

WP1 implemented a rich portfolio of generic crowdsourcing services, which can be configured and invoked in specific scenarios, taking into account more refined technical or application-related requirements. In this work package, we will integrate them with their machine-driven counterparts in form of lightweight APIs, creating a set of powerful services that can be added to state-of-the-art data value chain technology to improve their data governance capabilities. The resulting service ecosystem will cover the entire lifecycle of linked data management, from data harvesting to inspection and analysis, maintenance, and interlinking and will benefit both data publishers and consumers. It will be integrated into the Linked Data Stack infrastructure as free software ready to be adopted in further application scenarios and systems.

Description of work

Task 2.1 Multilingual data harvesting (M7-M24; Lead: INFAI; Participants: SWC, ONTOS) conTEXT¹⁰³ allows to semantically analyze text corpora (such as blogs, RSS/Atom feeds, Twitter etc.) and visualize the results. It is based on NLP frameworks like FOX, OntosMiner, and Eventos to detect entities and their relations in text. We will integrate the crowdsourcing services developed in WP1 to gather feedback on the performance of various components, e.g., named entity recognition and relation extraction. This feedback channel will be used to train the underlying NLP algorithms in order to assess and improve their precision and recall.

Task 2.2 Data inspection (M7-M24; Lead: INFAI; Participants: BROX, UNIST, SOTON) In this task we will provide a comprehensive data inspection solution that brings together the most important standard automatic assessment techniques for both schema and instance data with the crowdsourcing services delivered in WP1. We distinguish between the following basic (semi-automatic) approaches, each backed up by an existing service:

- Test-driven approaches assess the quality of instance data by executing manually written test cases against an RDF data set. RDFUnit is a test-driven data-debugging framework that can run automatically generated (based on a schema) and manually generated test cases against a data store. The developer defines and executes the unit tests and evaluates their results.
- Statistical approaches handle data sets that are less structured – such data sets are very common in real-world scenarios in which linked data has been generated from existing legacy information systems. The resulting linked data is often shallow (with respect to the schema), which makes the application of automatic inference less meaningful. CROCUS provides a semi-automatic statistical approach for instance-level error detection, which is agnostic of the underlying linked data knowledge base.
- Schema level approaches check if instance data adheres to its corresponding schema, following a closed-world assumption. ORE (ontology repair and enrichment) is a quality assessment tool for schemata. It shares several components with RDFUnit, i.e., in particular the same underlying DL-Learner framework and, therefore, requires the same validation techniques.

The aim is to expand RDFUnit, CROCUS and ORE into crowdsourcing in order to reduce the effort required to train each tool or interact with it. The task will mostly use microtask crowdsourcing, though in principle any approach can be applied to generate training data. The feedback of the crowd will be used for two purposes: to filter incorrect and confirm correct data entries, but also to establish a feedback loop to improve pre-defined or automatically derived rules. The second is relevant for statistical techniques such as the ones used in CROCUS.

Task 2.3 On-the-fly quality assurance (M13-M24; Lead: UNIST; Participants: INFAI) As already noted in WP1 classical crowdsourcing cannot compete with automatic approaches in terms of scale and performance. Even when highly parallelized, human computation processes yield delivery times that are on the average orders of magnitude worse than state-of-the-art technology though their accuracy is typically higher. This task aims will address these trade-offs for quality assurance scenarios with a strong real-time component. Existing quality

¹⁰³ <http://context.aksw.org>

assurance techniques from T2.2 will be enhanced with a real-time version of the crowdsourcing services delivered in T1.2 to support operations on data streams as opposed to static linked data and to react to rapidly changing environments with volatile requirements (which are common, for instance, in eCommerce.

Task 2.4 Link discovery (M7-M24; Lead: INF AI; Participants: SWC, ONTOS, SOTON) LIMES¹⁰⁴, SILK¹⁰⁵ and KnoFuss¹⁰⁶ are link discovery frameworks for the Web of Data that identify similar entities as well as duplicates in Web data sets. Declarative link discovery tools, in particular LIMES, can execute so-called link specifications, which contain heuristics for the similarities of entities in data sets. Those specifications can be either created manually or via machine-learning techniques. Human feedback is required to assess and maximize the precision and recall of these link specifications as well as resultant output.

Task 2.5 Usage analyses for data set maintenance (M7-M24; Lead: SOTON; Participants: SWC) The extent to which a data set is able to satisfy user queries is a strong indicator for its fitness of use, guiding data publishers in their data maintenance duties. In this task we will turn the USEWOD¹⁰⁷ tool chain¹⁰⁸ into an open source Web application to be used by external parties in business scenarios. An instance of the service will be openly deployed to be used by linked open data providers. This means that the analysis results will also be openly available (partially even as Linked Open Data) and hence directly contribute back to assess and assure the quality of the Web of Data in an evolutionary fashion. Some quality dimensions assessed by this usage analysis approach are critically affected by how well the performed queries fit to the data set. This allows for finding blind spots on the instance as well as the schema level of a data set, but it can also be exploited for polluting query logs with unreasonable, but well-formed queries. We will involve crowdsourcing to identify the most representative and meaningful queries for a data set.

Deliverables

All deliverables will include software binaries and documentation. The M24 version will consider the results of lab tests in the evaluation (see WPs 4 and 5).

- D2.1 Multilingual data harvesting services v1 (M12; Lead: INF AI)
- D2.2 Data inspection services v1 (M12; Lead: INF AI)
- D2.3 Link discovery services v1 (M12; Lead: SWC)
- D2.4 Data maintenance services v1 (M12; Lead: SOTON)
- D2.5 On-the-fly quality assurance tools and report on SLAs (M24; Lead: UNISTER)
- D2.6 Multilingual data harvesting services v2 (M24; Lead: INF AI)
- D2.7 Data inspection services v2 (M24; Lead: BROX)
- D2.8 Link discovery services v2 (M12; Lead: INF AI)
- D2.9 Data maintenance services v2 (M12; Lead: SOTON)

Work package description – WP3

Work package number	WP3		Start date or starting event				M7
Work package title	Quality-minded linked data tool suites						
Participant number	1	2	3	4	5	6	7
Participant short name	SWC	BROX	ONTOS	REEEP	UNIST	INF AI	SOTON
PM per participant	22	21	21	0	0	9	9

Objectives

The services developed in WP1 and WP2 are the functional base to extend existing data publishing and management tools in order to establish data quality assessment and repair before and after each individual step in the data value chain as a generic, scalable, accurate and cost-effective support activity. The outcome of this work package will be releases of three tool suites empowered with different types of crowdsourcing-based data quality assessment and repair capabilities, usable in a variety of application scenarios and vertical sectors. The results will be made available freely when possible (OntoWiki, PoolParty) or under commercial license alternatively (OntosLDIW). All open components will be made available via the Linked Data Stack. They are the technical backbone of the vertical data value chains realized in WPs 4 and 5, which will also be used to test the new

¹⁰⁴ <https://github.com/GeoKnow/LIMES-Service>

¹⁰⁵ <http://silk.wb3g.de>

¹⁰⁶ <http://technologies.kmi.open.ac.uk/knofuss/>

¹⁰⁷ <http://usewod.org>

¹⁰⁸ Markus Luczak-Rösch, Usage-dependent maintenance of structured Web data sets, PhD Thesis, Freie Universität Berlin, http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000096138

functionality. The relationship between WPs 2 and 3 can be explained as follows: WP2 deals with finer-granular, fundamental data management services and integrates crowdsourced and automatic assessment and repair methods; by contrast, WP3 operates at the level of more complex tool suites, which have a strong process component and incorporate various activities in the linked data management lifecycle.

Task 3.1 Crowdsourcing-enabled authoring via OntoWiki (M7-M24; Lead: BROX; Participants: INFAl)

In this task, we will extend OntoWiki with a data cleansing component based on services developed in WP1. OntoWiki is a knowledge engineering tool and development platform, which was applied in various scenarios such as requirements engineering, authoring of historical information, as well as authoring large taxonomies and data in the automotive industry. OntoWiki is also part of the commercial eccenca Linked Data Suite that is developed by BROX. Quality assessment and repair is essential, in particular in distributed authoring scenarios. The development will include microtask crowdsourcing, gamification and open challenges.

Task 3.2 Crowdsourcing-enabled OntosLDIW (M7-M24; Lead: ONTOS; Participants:-) Crowdsourced data cleansing and repair will be applied to named entity disambiguation (NES), co-reference resolution (CR), unique identifier (UID) attribution, as well as merging and interlinking in OntosLDIW. The integration of the new features will occur at two stages: (1) we will create HITs based on the output of OntosMiner (RDF/XML) to ask workers to disambiguate named entities; and then (2) develop an user interface that will interact with the Identifier Knowledge Base (IKB) component of OntosLDIW, assigning HIT jobs to solve the problem of UID, merging and linkage (e.g., owl:sameAs) to other relevant data sets. Last, but not least we will use microtask crowdsourcing to create a large set of alternative entity names in different languages. Since an entity can be called in a very different manners and languages (e.g., “Madonna” or “La reina del pop”@es), this tool will use machine translation and machine learning to determine label suggestions for the crowd to validate. Moreover, users will be able to identify the uniqueness of a named entity, and in case of no uniqueness, pass to a stage of merging, which also has to be crowd validated. The targeted crowds will be preferred to be multilingual experts. We will experiment with multiple crowdsourcing workflows, in particular when generating open text for labeling and translation (see T1.2).

Task 3.3 Enhanced quality assessment and repair mechanisms in PoolParty (M7-M24; Lead: SWC; Participants:-) PoolParty Thesaurus Server (PPT) is an advanced software platform to manage enterprise metadata (taxonomies, thesauri, controlled vocabularies) and linked data. PPT’s metadata management is based on RDF and SKOS and exploits text mining and linked data mapping technologies in addition to manual editing. Based on our existing QSKOS plug-in, we will enhance the functionality of PPT according to the quality needs elaborated in REEEP business scenario:

- **Taxonomy enhancement:** Using a crowdsourced approach for enhancing already mature and grown taxonomies with input from new sources (like analysed content streams, social media streams and social tagging) where the quality of this crowdsourced concepts is ensured by mechanisms developed in WP1. Resulting quality indicators will guide the thesaurus manager in choosing and aligning concepts in the process of taxonomy enhancement.
- **Taxonomy management:** Involve domain expert groups into quality and repair mechanisms to support the quality of relevance, structure and consistency of the growth and fine-tune of taxonomies.
- **Formal quality:** The check for inconsistencies and missing linkage within a mature corporate thesaurus in using rule based quality and validation concepts.
- **Link quality:** The linkage with external vocabularies will be validated and quality proofed by using (not only) crowdsourcing e.g., in running a game based approach asking for the likeliness and relevance of links.

The results, together with metadata about their production process and quality (see T1.3) will be returned to the thesaurus management environment via an interactive dialogue. They will be validated by the user and integrated with the rest of the thesaurus.

Deliverables

All deliverables will include software binaries and documentation. The M24 version will consider the results of lab tests in the evaluation (see WPs 4 and 5).

D3.1 OntoWiki v1 (M12; Lead: BROX)

D3.2 OntosLDIW v1 (M12; Lead: ONTOS)

D3.3 PoolParty v1 (M12; Lead: SWC)

D3.4 OntoWiki v2 (M24; Lead: BROX)

D3.5 OntosLDIW v2 (M24; Lead: ONTOS)

D3.6 PoolParty v2 (M24; Lead: SWC)

Work package description – WP4

Work package number	WP4		Start date or starting event				M4
Work package title	Data value chain for the energy sector						
Participant number	1	2	3	4	5	6	7
Participant short name	SWC	BROX	ONTOS	REEEP	UNIST	INFAI	SOTON
PM per participant	30	0	3	32	0	0	5

Objectives

This work package will realize a data value chain in the clean energy sector with enhanced data quality assessment and repair capabilities. It will build on a clear business case and existing industry partners and projects between REEEP and SWC. SWC supplies Linked Open Data technology for REEEP's Reegle Content Pool, the Reegle Tagging API and Reegle Thesaurus services to implement an intelligent information management solution on topics related to renewable energy, energy efficiency and climate compatible development. The two main outputs of this work package are 1) an enriched Reegle Content Pool with high-quality compilations, which is quality assured based on continuous improvement mechanisms built into the supporting tools including trends analysis of content and crowdsourcing-enabled services delivered by WP1 and WP2; 2) renewable energy, energy efficiency and climate related on-demand structured statistical data, leveraging and multiplying the value of the improved Content Pool; 3) an enhanced Reegle Thesaurus based on crowdsourcing outputs from WP1 and WP2, completing and continuously verifying translations of concepts in the thesaurus into French, Spanish, Portuguese and German (as default language is English); 4) Improved Reegle Tagging API results based on algorithm and user feedback continuous improvement implementation ; 5) Information Management Consultancy Services to help all current organizations understand how to best take advantage of all these developments as well as to market, disseminate and communicate these advantages to future customers for the REEEP portfolio and the international agenda to combat climate change. The curated data will be published as Linked Open Data wherever possible, thereby creating a sustainable and current solution for all organizations operating in this market, including those already using REEEP's Reegle services for their information management. This solution adopts the data value chain and creates a self-sustained tool and customer base to utilize it based on its continuous improvement and quality assurance mechanisms in place.

Task 4.1 Data collection and enrichment (M4-M6; Lead: REEEP; Participants: SWC) This task is concerned with the harvesting of reputable, relevant open data from international organisations relevant to renewable energy, energy efficiency and climate compatible development. This includes engaging with stakeholders through REEEP's network as well as automatic harvesting methods. This task will also include the publication of primary sources as linked data and the enrichment and interlinking of the resulting machine-understandable data sets.

Task 4.2 Technical development (M4-M27; Lead: SWC; Participants: REEEP, SOTON) We will integrate the newly harvested data with existing REEEP and organizational data in the Reegle Content Pool according to a standardized process consisting of the following steps: (1) scrapping; (2) quality assessment and repair; (3) interlinking of businesses and actors catalogues; (4) revise algorithms for the Reegle Tagging API; (5) create technical options for organizations implementing the Reegle Tagging API to select thematic areas of the Reegle thesaurus; and finally (6) enable organizations to seamlessly push content to the Reegle Content Pool (automatically or manually depending on their configuration). The resulting much improved Reegle Content Pool is set up for continuous harvesting, scrapping, integration, analytics, and dissemination to external and internal Linked Open Data solutions. Additionally automated tagging of all new content will be implemented in order to make new themes available for further Reegle Thesaurus development. We will make use of crowdsourcing to associate trending topics on Twitter to themes in the content pool as a basis for the generation of on-demand, tailored reports, as well as for the validation of steps (1) and (3) in the process outlined above. We will make use of different crowdsourcing approaches, including gamification for organization-internal purposes, an open challenge for energy-aware data enthusiasts, and paid microtasks.

Task 4.3 Market validation (M13-M30; Lead: REEEP; Participants: SWC, SOTON) We will release a new version of the Reegle Content Pool, the Reegle Tagging API and the Reegle Thesaurus to REEEP's international network of existing and potential customers, showcasing the new functionality solving user needs in terms of different thematic choices of the Reegle Tagging API's Thesaurus, automatically cleaned Reegle Content Pool for broken and duplicate links as well as automated harvesting of their own knowledge management which they require such solutions for in order to save time and resources which are precious in light of their clean energy and climate change mandates. In addition, members of the network will be approached to contribute further to development of the Reegle Thesaurus to additional thematic areas and to use the Reegle Tagging API to tag their documents as well as other assets available in the new Reegle Content Pool. We will produce new reports to

display the robustness of the content and emphasize the value of crowdsourcing, comparing the latter with existing benchmarks. We understand the technology and associated data marketplace for renewable energy, energy efficiency and climate compatible development as an enabler of Social Entrepreneurship. Novel products and services in the clean energy sector require curated data and sophisticated data management capabilities, as well as ad-hoc reporting options to respond to the ever changing knowledge canvas of clean energy and climate compatible development and the policy decisions associated with these challenges.

Deliverables

D4.3 includes deployed software as well as a technical specification of the solution detailing the business case as per the demand already identified, consultancy with existing users, developed as a quality insured Linked Open Data source with continuous improvement and accuracy as key priorities.

D4.4. includes test plans, user acceptance testing, results reports, adjustments, improvements, integration of end-user feedback including new adopters of the tool, and existing users of the tool.

D4.1 Product development specification (M6; Lead: REEEP)

D4.2 Data harvesting for the energy data value chain (M6; Lead: REEEP)

D4.3 Data value chain in the energy sector v1(M18; Lead: SWC)

D4.4 Testing and evaluation report (M18: Lead: SWC)

D4.5 Data value chain in the energy sector v2 (M27; Lead: SWC)

D4.6 Market validation report (M30; Lead: REEEP)

Work package description – WP5

Work package number	WP5		Start date or starting event				M4
Work package title	Data value chain for the eCommerce sector						
Participant number	1	2	3	4	5	6	7
Participant short name	SWC	BROX	ONTOS	REEEP	UNIST	INFAL	SOTON
PM per participant	0	20	7	0	30	8	5

Objectives

In this work package we will use the results and services defined and implemented in WP1, WP2 and WP3 and apply them to enhance Unister’s eCommerce platform. The main objective is to define a data hub containing pricing informatin. The main outcomes of the work package are twofold: 1) a service providing access to a pool of general-purpose eCommerce content; the underlying knowledge base of the service uses data in different languages harvested from public and enterprise sources, and curated using services delivered by WPs 1 and 2; 2) different parts of the knowledge base can be requested on demand by querying for different quality levels; 3) the implemented services are applied to the travel domain, which is a major eCommerce vertical that will also be used for market validation purposes.

Task 5.1 Data collection and enrichment (M4-M6; Lead: UNIST; Participants: BROX, INFAL) This task is concerned with the harvesting of reputable, open data from the linked open data cloud relevant to the travel domain. A focus will be put on geo-spatial data (e.g., places, POIs and hotels), as they are core to any online travel eCommerce experience. In addition enterprise data sets of Unister and external data sets of Unister’s business partners will be taken into account and a quantitative case study will be created based on real-world user data (in anonymized form). The data will be translated into Linked Open Data and interlinked to major data hubs in the global Web of Data graph.

Task 5.2 Technical development (M4-M27; Lead: UNIST; Participants: BROX, ONTOS, INFAL, SOTON)

The data value chain for this eCommerce scenario looks as follows; (1) retrieve data sets; (2) link products to user intentions and locations as well as different instances of the same data; (3) apply the algorithms for data analysis (including Web content and sentiment analysis); (4) implement interfaces for adding data to the data store of the service; (5) implement interfaces for fetching data from the service; (6) implement interfaces for pushing repaired data by knowledge carrier to the service (e.g., by the quality assurance tools). After each step additionally quality assessment and annotations of computed quality metrics will be applied. We will use primarily social-network-based crowdsourcing (engagement, gamification) paired with paid microtasks for the tasks that appeal less to end-customers.

Task 5.3 Market validation (M13-M30; Lead: UNIST; Participants: BROX, ONTOS, SOTON) We will rely on real-world interactions to simulate actual usage of the computed data set via the service delivered in T5.2. We will resort to business experts to uncover problems with the data sets and trigger requests for improving the data quality. We will hence validate the processes and the impact of data curation results in periodic reports and prove the robustness of the processes by comparing the newly developed service against the

state-of-the-art solution which uses very basic quality assurance and repair methods.

Deliverables

D5.3 includes deployed software as well as a technical specification of the solution detailing the business case in detail as per the demand already identified, consultancy with existing users, developed as a quality insured Linked Open Data source with continuous improvement and accuracy as key priorities.

D5.4. includes test plans, user acceptance testing, results reports, adjustments, improvements, integration of end-user feedback including new adopters of the tool, and existing users of the tool.

D5.1 Product development specification (M6; Lead: UNIST)

D5.2 Data harvesting for the eCommerce data value chain (M6; Lead: UNIST)

D5.3 Data value chain in the eCommerce sector v1(M18; Lead: UNIST)

D5.4 Testing and evaluation report (M18: Lead: BROX)

D5.5 Data value chain in the eCommerce sector v2 (M27; Lead: UNIST)

D5.6 Market validation report (M30; Lead: UNIST)

Work package description – WP6

Work package number	WP6		Start date or starting event				M1
Work package title	Impact creation						
Participant number	1	2	3	4	5	6	7
Participant short name	SWC	BROX	ONTOS	REEEP	UNIST	INFAI	SOTON
PM per participant	6	6	7	8	6	5	5

Objectives

The general aim of this work package is to establish a worldwide focal point for academic and industry parties interested in contributing to or taking advantage of the crowdsourcing-driven data cleansing and repair solution delivered by the project. More information is available in Section 2.2

Description of work

Task 6.1 Dissemination and networking (M1-M30; Lead: ONTOS; Participants: all) We will promote the project on the Web, and through social media. The applications in WPs 4 and 5 will be demonstrated at core industrial events. Key R&D results will be published in conferences and journals in the linked data, crowdsourcing, and HCI communities. Open challenges and public service deployments will bring in additional visibility among data publishers and developers.

Task 6.2 Exploitation (M1-M30; Lead: SWC; Participants: all) The application scenarios and the crowdsourcing services are the focus of the exploitation. The major outcomes of the task will be the exploitation plan, which will include a list of opportunities and a detailed analysis of benefit and impact, and the final report providing details about the technology transfer and commercialization potential foreseen.

Task 6.3 Standardization (M1-M30; Lead; SOTON; Participants: all) In this task we will promote the vocabulary developed in T1.3 to W3C as well as grassroots initiatives and crowdsourcing technology providers.

Deliverables

D6.1 Project fact sheet, press releases and online presence (M1; Lead: ONTOS)

D6.2 Exploitation plan (M6; Lead: BROX)

D6.3 Dissemination and networking report v1 (M18; Lead: ONTOS)

D6.4 Dissemination and networking report v2 (M30; Lead: ONTOS)

D6.5 Exploitation report (M30; Lead: SWC)

D6.6 Standardization report (M30; Lead: SOTON)

Work package description – WP7

Work package number	WP7		Start date or starting event				M1
Work package title	Project management						
Participant number	1	2	3	4	5	6	7
Participant short name	SWC	BROX	ONTOS	REEEP	UNIST	INFAI	SOTON
PM per participant	18	1	1	1	1	1	1
Objectives							
<p>This work package will ensure the smooth management of the project, by making sure that the structures and processes defined in Section 3.2 and in contractual agreements are followed. This includes the strategic, technical, as well as the administrative and financial measures to ensure:</p> <ul style="list-style-type: none"> • that the project remains on course, • that it is effectively and correctly managed financially, • that its progress and status are efficiently and effectively monitored, • that the required reporting is prepared and delivered in a timely manner, • that all quality aspects of the project are fully and correctly addressed, • that support for the infrastructure supporting the Web-based facilities to be used for dissemination and central intra-project communication and cooperation is provided. <p>The project management will entail strategically, project-wide as well as day-to-day central management and coordination activities. The several different management boards which will be established in the consortium will be responsible for decisions and activities of different scope and level according to their function.</p>							
Task 7.1 Administrative and financial management (M1-M30; Lead: SWC) This task is responsible for setting up the initial project contract and to manage financial and administrative issues over the whole project duration.							
Task 7.2 Project coordination (M1-M30; Lead: SWC) This task subsumes the general project coordination including communication with the European Commission, management of the various boards defined in Section 3.2 of the proposal, and the overall responsibility of the quality management of the project and of its results.							
Task 7.3 Research data management (M4-M30; Lead: SWC) This task is about all measures to be taken to participate in the “pilot action on open access to research data“ from formulating the initial Data Management plan to define scope, standards, metadata and sharing technologies for open up research data generated by the project							
Deliverables							
<i>Most deliverables are contractual deliverables specified in the Grant Agreement.</i>							
D7.1 Research data management plan (M3; Lead: SWC)							

3.1.3.2 Work package list

WP no	Work package title	Lead participant no	Lead participant short name	Person months	Start month	End month
WP1	Crowdsourced data quality services	7	SOTON	41	M1	M18
WP2	Hybrid data quality services	6	INFAI	73	M7	M24
WP3	Quality-minded tool suites	2	BROX	82	M7	M24
WP4	Data value chain in the energy sector	4	REEEP	70	M4	M30
WP5	Data value chain in the eCommerce sector	5	UNIST	70	M4	M30
WP6	Impact creation	1	SWC	43	M1	M30
WP7	Project management	1	SWC	24	M1	M30
			TOTAL	403		

3.1.3.3 *List of deliverables*

Del. (no)	Deliverable name	WP no	Short name of lead part.	Type	Diss. level	Delivery date
D1.1	Crowdsourcing design	WP1	SOTON	OTHER	PU	M3
D1.2	Crowdsourcing services v1	WP1	SOTON	OTHER	PU	M6
D1.3	Methods for workflow, task, and time management	WP1	SOTON	R	PU	M12
D1.4	Crowdsourcing services v2	WP1	INFAI	OTHER	PU	M12
D1.5	Crowdsourcing vocabulary and licensing	WP1	SOTON	OTHER	PU	M12
D1.6	Public endpoints and deployment	WP1	INFAI	OTHER	PU	M18
D2.1	Multilingual data harvesting services v1	WP2	INFAI	OTHER	PU	M12
D2.2	Data inspection services v1	WP2	INFAI	OTHER	PU	M12
D2.3	Link discovery services v1	WP2	SWC	OTHER	PU	M12
D2.4	Data maintenance services v1	WP2	SOTON	OTHER	PU	M12
D2.5	On-the-fly quality assurance tools and report on SLAs	WP2	UNIST	R + OTHER	PU	M24
D2.6	Multilingual data harvesting services v2	WP2	INFAI	OTHER	PU	M24
D2.7	Data inspection services v2	WP2	BROX	OTHER	PU	M24
D2.8	Link discovery services v2	WP2	INFAI	OTHER	PU	M24
D2.9	Data maintenance services v2	WP2	SOTON	OTHER	PU	M24
D3.1	OntoWiki v1	WP3	BROX	OTHER	PU	M12
D3.2	OntosLDIW v1	WP3	ONTOS	OTHER	CO	M12
D3.3	PoolParty v1	WP3	SWC	OTHER	PU	M12
D3.4	OntoWiki v2	WP3	BROX	OTHER	PU	M24
D3.5	OntosLDIW v2	WP3	ONTOS	OTHER	CO	M24
D3.6	PoolParty v2	WP3	SWC	OTHER	PU	M24
D4.1	Product development specification	WP4	REEEP	R	PU	M6
D4.2	Data harvesting for the energy data value chain	WP4	REEEP	OTHER	PU	M6
D4.3	Data value chain in the energy sector v1	WP4	SWC	OTHER	PU	M18
D4.4	Testing and evaluation report	WP4	SWC	R	PU	M18
D4.5	Data value chain in the energy sector v2	WP4	SWC	OTHER	PU	M27
D4.6	Market validation report	WP4	REEEP	R	PU	M30
D5.1	Product development specification	WP5	UNIST	R	PU	M6
D5.2	Data harvesting for the energy data value chain	WP5	UNIST	OTHER	PU	M6
D5.3	Data value chain in the eCommerce sector v1	WP5	UNIST	OTHER	PU	M18
D5.4	Testing and evaluation report	WP5	BROX	R	PU	M18
D5.5	Data value chain in the eCommerce sector v2	WP5	UNIST	OTHER	PU	M27
D5.6	Market validation report	WP5	UNIST	R	PU	M30
D6.1	Project fact sheet, press releases and online presence	WP6	ONTOS	DEC	PU	M1
D6.2	Exploitation plan	WP6	BROX	R	CO	M6
D6.3	Dissemination and networking report v1	WP6	ONTOS	R	PU	M18
D6.4	Dissemination and networking report v2	WP6	ONTOS	R	PU	M30
D6.5	Exploitation report	WP6	SWC	R	CO	M30
D6.6	Standardization report	WP6	SOTON	R	PU	M30
D7.1	Research data management plan	WP7	SWC	R	PU	M3

3.1.4 Graphical representation of component dependencies (Pert diagram)

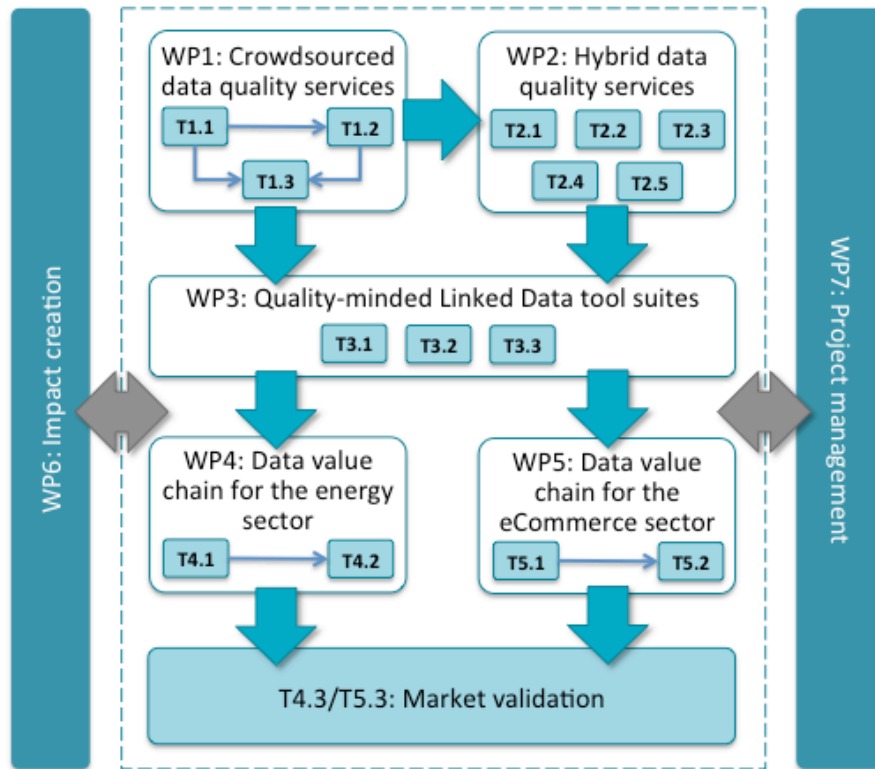


Figure 3.1.2 Dependencies of tasks in QROWD

The QROWD work plan is set up around the extension of three widely-applied linked data tool suites, which will be extended to facilitate the data quality support activity more easily. For this purpose the tool development in WP3 will exploit the generic data quality assessment services developed in WP1 and WP2. The adapted tools will then be tested and evaluated in two relevant business cases, representing a significant share of today’s digital economy in Europe. The final task of WP4 and WP5 is the market validation step, which is meant to allow insight about the particular tools applied but also the underlying generic approaches. The work packages on impact creation and project management embrace and support the technical core throughout the entire project runtime.

3.2 Management structure and procedures

The project management strategy of QROWD is tailored to the specific scope and aims of the project, and in particular to the number and type of contractual partners and potential collaborators. The strategy will ensure that all of the key objectives of the project are achieved within time, cost, and resource constraints. It will use tried-and-tested project management structures, procedures and tools, as well as technical support that leverage the project management experience of the organizations involved and the project coordinator SWC in particular. The project is governed by the contracts with the two stakeholder bodies: the European commission (through the EC Grant Agreement) and the General Assembly of QROWD partners (through the Consortium Agreement).

3.2.1 QROWD structures

In the remainder of this section we give a brief overview of the different bodies and roles within the management structure of the project, and briefly sketch the most important procedures.

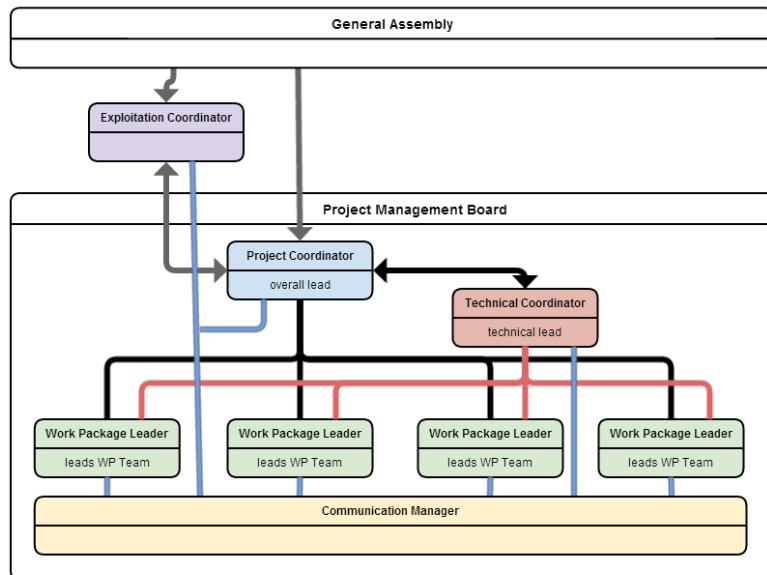


Figure 3.2.1 QROWD management structure

3.2.1.1 Project Coordinator (PC)

The PC has the overall responsibility for intermediation between the consortium and the European Commission, as well as for the financial and contractual obligations defined in the Grant Agreement and last but not least the coordination between the 7 work packages. The PC of QROWD will be managed by SWC, in person by Martin Kaltenböck (SWC), a project coordinator and CFO who has a long time experience and track record of successful collaborative research projects as well as mid- and large scale industry projects in which he played an active role as project coordinator.

The PC is further on responsible for the following activities: the production and timely submission of reports to the EC; the implementation and continuous improvement of adequate project management procedures such as quality assurance and risk assessment; and the set up and maintenance of the necessary infrastructure for intra-project communication and management.

3.2.1.2 Technical Coordinator (TC)

The TC has to facilitate technical coordination and integration between work packages. The TC will ensure the technical and architectural consistency of software and research activities throughout the project by

- Providing ongoing technical overview and direction for the project;
- Aligning the technical requirements across the work packages;
- Managing interdependencies between work packages;
- Defining interfaces between work packages software components;
- Facilitating technical communication between project participants;
- Taking care that all development results are tested (quality management) and documented.

Recommendations for software development practices, technical interfaces and project architecture will be reported from the Technical Coordinator to the PMB enforced by the WPL (see below).

3.2.1.3 Work Package Leaders (WPL)

Each work package has a dedicated leader (organization and person), who is responsible for coordination and is assigned by the leading partner of the respective WP. She supervises the development of a detailed work plan, and ensures that all appointed deliverables meet the project objectives and are completed in due time.

3.2.1.4 Project Management Board (PMB)

The PMB is in charge of the execution and overall management of the Development and Innovation activities of the project. It coordinates the integration of all activities in the work plan; monitors the progress of the WPs; proposes changes in work sharing, budget and participants to the coordinator (PC) and reviews and monitors technical progress. Furthermore the PMB manages gender equality aspects, as described in section 1.3. The PMB consists of the leaders of the WPs, and the project coordinator, who chairs the board, communicates at least monthly (monthly calls), and meets at least once every six months (plenary and PMB meetings).

3.2.1.5 *General Assembly (GA)*

The GA serves as supervising body for the definition and review of the overall project progress and acts as a forum for high-level decision-making and as an action-wide discussion platform for the approval of budgets and work plans. Furthermore the GA monitors gender equality aspects. It consists of one delegate of each of the partners (C-level), meets at least once within the project duration and is chaired by the Project Coordinator.

3.2.1.6 *Exploitation Coordinator (ECO)*

ECO takes responsibilities for a coordinated and effective “way to market” of the business scenarios pioneered with the projects technologies, as well as the technology exploitation as such. As the targeted market is young exploitation activity including also market stimulation and awareness raising. Therefore, the ECO has to take care that activities are carried out, which:

- build awareness of open data business in the respective communities
- built awareness about the importance of data quality mechanisms in data management
- provide best practises and business models for open data (quality) management
- attract open data Entrepreneurs to invest in open data quality services
- disseminate information and facts about the business impact of QROWD technologies and approaches

3.2.1.7 *Communication Manager (CM)*

All internal and external communication, which is not handled by other project management bodies (e.g., management reports by the PC, documentation by the TC, marketing activities by the ECO) will be coordinated by the CM. This involves traditional public relation work, the maintenance of the external Internet services (Web site, social media, etc.) as well as the cross-marketing activities with the PR channels of the QROWD partners.

3.2.2 *Decision procedures*

The primary mechanism for decision-making throughout all groups within the project will be by consensus (defined as a lack of sustained opposition to a decision).

3.2.2.1 *Consortium wide decisions*

The partners will be bound by a formal Consortium Agreement that will be signed before entering the Grant Agreement with the EC, and in which their roles, responsibilities and mutual obligations will be defined both for the project life and vis-à-vis intellectual property. The Consortium Agreement will also address the issues of decision-making and conflict resolution.

Management Board meetings will be organized every 6 months or as necessary if important decisions need to be taken as a result of unforeseen circumstances. Each Consortium member (as represented in the PMB) will hold one vote regardless of its share of project budget. Extraordinary decisions shall be taken at a qualified majority of 75% of votes and of members present or represented by proxy at a meeting, provided always that any partner whose scope of work, time for performance, costs or liabilities are changed or whose information is to be published, may veto such decisions. By exception for dealing with a defaulting partner, decisions shall be taken unanimously by all of the non-defaulting partners.

3.2.2.2 *Decisions within work packages*

Within a work package, decisions should be based on consensus among the WP team. Whenever consensus cannot be reached on a given issue, the WP team will ask for the arbitration of the Technical Coordinator. If the WP team rejects the result from this process, the issue will be put on the agenda of the next PMB meeting.

3.2.2.3 *Conflict resolution*

Where consensus cannot be reached, it is essential that processes should be available to escalate disagreements. The procedures outlined below are defined in full in the consortium agreement including rules for convening a meeting, definition of a quorum and voting. The potential for some conflict in a complex Integrating Project must be regarded as high because it involves individuals from different backgrounds and organisational cultures working together to complete a complex set of tasks. Day-to-day conflicts may relate to differences in priorities, resource allocation, technology choices, ways of working, or expectations of results. The conflict resolution mechanism described below reflects the overall project management structure and philosophy of devolved responsibility. The primary aims are to minimise the impact of any disagreement and to localise its effect; and to ensure the speediest possible resolution of disagreements.

Disputes localized within a work package

Where there is sustained disagreement within a work package, which the work packages leader is unable to mediate, the Project Coordinator will be invited to mediate and in the absence of consensus, the Project management board shall be invited to mediate.

Disputes between work packages

Conflicts between work packages shall, in the first instance, be mediated by the Project Coordinator through the Project management board. If the Project management board is unable to reach consensus, the disagreement shall be referred to the General Assembly. Conflicts between a work package and the Project Coordinator shall be managed in the same way as disputes between work packages, except that the project management board shall elect a chairperson to temporarily replace the Project Coordinator in that role.

Disputes between institutions

Conflicts between the Project Coordinator and a Consortium Member, or between Consortium Members, shall be referred directly to the General Assembly (although the Project Coordinator may first be invited to mediate disputes between Consortium Members). The General Assembly shall be the final point of decision for internal resolution of all conflicts. Within its own areas of competence, and enacting as the point of appeal from decisions of the Project management board, it shall always seek to make decisions by consensus. In the event that a conflict between institutions cannot be resolved internally, Consortium Members shall retain the option to pursue conflict resolution through external channels, as described in full in the Consortium Agreement.

3.2.3 Quality control

The quality assurance of contractual deliverables starts four weeks before the deliverable submission. All deliverables have an assigned reviewer from the project team who is not involved in the authoring of the deliverable. The feedback by the reviewer is documented and implemented under the supervision of the WPL by the deliverable leader and contributors. In a subsequent step, the deliverable undergoes an additional release check by the PC. This second stage of the quality assurance procedure might call for additional quality improvements by the deliverable authors. The guidelines and procedures for quality and risk management will describe the activities and resources necessary to ensure that the quality requirements of the project are met. It will define quality standards based on ISO 9000:2000 principles, quality requirements, quality assurance methods, quality assurance activities and configuration management. It will also define policies for identifying threats on the project and for implementing corrective actions as well as the policy for reviewing deliverables and public documents before their issuance.

3.2.4 Planning and reporting

Reporting will be ensured by the Project Coordinator and will adhere to the practices of the EC Project Office. Periodic activity reports (usually quarterly) will be prepared containing overview of the activities carried out in the respective period, including the implementation of outreach goals, a description of progress towards the objectives and the milestones, and deliverables foreseen, the identification of problems encountered and corrective actions taken. Furthermore, we will develop and regularly update a plan for using, disseminating and exploiting knowledge. This plan shall be included in the periodic activity reports. The PC is also responsible to organize and lead the review meeting with the EC project officer and the project consortium that will take place three times (annually) or two times (mid of project and end of project) depending on EC rules. Finally, the project management (PC) will be in charge of producing periodic management reports, which will include:

- A justification of the resources deployed by each beneficiary, linking them to activities implemented and justifying their necessity;
- The financial statements provided by each beneficiary for that period;
- A summary financial report consolidating the claimed costs of all the beneficiaries in an aggregate form;
- A report on the distribution between beneficiaries made during that period of the Community financial contribution.

Table 3.2a List of milestones

Milestone number	Milestone name	Related work package(s)	Estimated date	Means of verification
MS1	Basic concept and specification available	WP1, WP4, WP5, WP6, WP7	M6	Completed deliverables D1.1, D1.2, D4.1, D4.2, D5.1, D5.2, D6.1, D6.2, D7.1 Crowdsourcing services are available as early prototypes Product layouts are defined for all business cases
MS2	First trial services	WP1, WP2, WP3	M12	Completed deliverables D1.3, D1.4, D1.5, D2.1, D2.2, D2.3, D2.4, D3.1, D3.2, D3.3 First integration in partners product suites
MS3	Refinement specification and validation	WP1, WP4, WP5, WP6	M18	Completed deliverables D1.6, D4.3, D4.4, D5.3, D5.4, D6.3 First integration and validation of tool bundles in business scenarios
MS4	Second trial services	WP2, WP3	M24	Completed deliverables D2.5, D2.6, D2.7, D2.8, D.2.9, D3.4, D3.5, D3.6 Crowdsourcing services have been deployed and are available for public use.
MS5	Tools and plugin final release and documentation ready	WP4, WP5, WP6	M30	Completed deliverables D4.5, D4.5, D4.6, D5.5, D5.6, D6.4, D6.5, D6.6 Final integration and validation of tool bundles in business scenarios Overall market validation and future potential analysis Documentation and release of public documents and free software (where applicable)

3.2.5 Risk management

Risk management is the process of deciding what, if anything should be done with a risk. It answers two key questions:

- Who owns the risk (responsibility)?
- What can/should be done (scope and actions)?

The range of response actions for QROWD is as follows:

- Delegate: Risk is internal to project.
- Research: Means investigating the risk until you know enough to be able to decide what to do. Research can range from making a few phone calls to conducting a detailed review of the risk drivers.
- Transfer: Risk is external to QROWD. Resources and knowledge outside of QROWD are better able to manage the risk. Transfer implies the ultimate accountability, responsibility and authority to expend resources. Transfer requires acceptance of the risk by the receiving party.
- Accept: Do nothing. Handle the risk as an issue if it occurs. No further resources are expended in managing the risk. These are usually risks that are not significant enough to justify any expenditure. Acceptance will include developing a contingency plan to execute should a risk occur. The most usual acceptance response is to establish a risk contingency, or reserve, including amounts of time, money or resources to use should the risk become a problem.
- Mitigate: Eliminate or reduce the risk by reducing the impact, reducing the probability, or shifting the timeframe when action must be taken.
- Watch: Monitor the risks and their attributes for early warning of critical changes in impact, probability, timeframe or other aspects. Watched risks are usually those for which existing conditions are not favorable for taking action (i.e., wait for improved conditions); or the potential for significant impact exists but the likelihood is low.

Table 3.1b Critical risks for implementation

Description of risk	WP(s) in-volved	Proposed risk-mitigation measures
Partner leaves consortium	All	All partners are fully committed to QROWD and have a vested interest in participating in the project to expand their product portfolio with effective means to deal with quality assessment and repair in the Linked Open Data Cloud. As it can be seen from Table 3.3.1 in Section 3.3 key expertise dimensions are covered without exception by more than one partner. In the unlikely case of a partner leaving the consortium, the remaining partners have the possibility to select a new partner with a similar profile; though highly undesirable, this replacement is feasible, given the large business network of the consortium and a developing field of technology providers and adopters of Linked Open Data. Preparation and recruiting by PMB and PC with a final vote coming from GA.
Personnel planning and management issues	All	Our partners are fully responsible for their own staffing and recruiting; this distributes and balances the risk for the overall consortium. We will further reduce this risk via careful monitoring and resource planning under the lead of the project management team. As mentioned earlier, the knowledge and expertise required to achieve the main goals of QROWD is covered by several partners. Additionally, the know-how generated in project will be available to new staff members or staff inheriting a task area through the project wiki.
Research, development or software components fail or provide limited functionality	WP1, WP2, WP3	Components will build upon in-depth expertise and proven and tested experimental settings available in the literature and extensively used by the partners in the past. From an implementation point of view services are straightforward to develop and will be integrated in existing, commercial products, which are extensible and mature. All partners, technology providers as well as universities have a great track record of software development and deployment. A switch between used and improved tools and components has to be prepared by the WPL and discussed and decided in the PMB.
Under-estimation of development time	All	Agile development processes support the continuous re-estimation of the required development time. Given an underestimation or a delay, the time plan as presented in Section 3.1.2 will be revised and necessary corrective actions will be taken. In the case of delay of a crucial component, partners have the resources to appoint additional back-up labour to the task.
Community initiatives fail to take up	WP1	The open challenges planned in QROWD will be centred on LOD sets that are central to many projects and developments. The need for curation and better interlinking is fully acknowledged by a great majority of the stakeholders, and efficient crowdsourcing methods as those applied in QROWD are enjoying immense popularity in the LOD community. Initial trials confirm this expectation, as they have been received very positively. Our public deployments will capitalize on these promising prospects and target the existing community of contributors around the data sets.
Microtask platform cannot be used as planned	WP1, WP2	The project will rely on different platforms for microtask crowdsourcing, reducing the risk of being too dependent on single providers. In addition, we will experiment with the usage of diverse types of crowds executing microtask-style work as showcased in the product lifecycle management case study.
Business opportunities become outdated	WP3, WP4, WP5, WP6	The business opportunities are tightly aligned with the development roadmap of the five industrial partners and reflect real problems in the use and reuse of Linked Open Data in critical adoption scenarios. The risk of these scenarios to become obsolete is rather improbable, as we discuss in Section 3.1.

3.3 Consortium as a whole

QROWD’s aims and objectives require a consortium consisting of organizations and individuals possessing complementary skills and technology know-how in the area of open data, and insight of the challenges practitioners face when attempting to use it in real-world applications. Also on a commercial level, the consortium is able to cover various parts and positions in the data value chain and the data lifecycle. As all commercial partners running business either on providing tools or providing content and services already in current activities, they have a good knowledge of threads which hinder a commercial exploitation of data on various positions. They can name these positions and provide solutions in providing tools and formulating business challenges.

3.3.1 Overall consortium

So in each of the two business areas, two business scenarios are adapted by totally 5 commercial partners – resulting in a bundle of business opportunities carried out by the project. In placing quality mechanisms in the centre of the project this efforts are exchangeable between a varieties of positions on different places at the lifecycle. In applying similar mechanisms of quality comprehensive experience with different forms of crowdsourcing it is crucial for the success of the project in order to devise effective methods and mechanisms that boost data quality, thus leading to a more purposeful exposure of newly published and existing data sets.

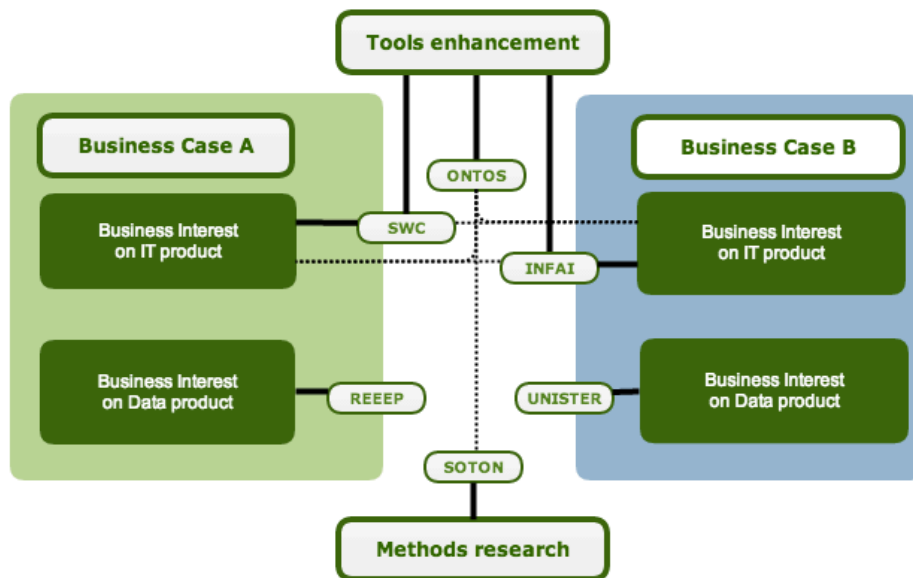


Figure 3.3.1 Contributions to individual partners

For the energy data value chain REEEP and SWC are partnering to get connected business interests for both partners realized under a common scenario. Whereby the technology partner (SWC) has specific interests in enhancing its ICT products, the data partner gets his interests on data products and services complied. A similar structured scenario with the partners Unister, BROX and Ontos will make the eCommerce data value chain. Tools and modules are exchanged between those two cases and the additional partner Ontos. INFAs and Southampton University’s role – as being the academic partner - is to bring state of the art technology and freshest research and development results into the development and innovation lifecycle.

3.3.2 Individual contributions

Software development expertise is essential in order to produce technology components that meet the robustness and performance constraints of typical application scenarios, and to integrate these components into existing (commercial) products and the vertical-sector deployments thereof. In-depth knowledge of human computer interaction and usability evaluation, as well as of the productive environments and application scenarios in which our software will be deployed is a pre-requisite for the design of crowdsourcing-enabled data cleansing components that meet the needs and expectations of end-users and fit well in their daily activities and workflows. Finally, ensuring that the results of the project reach out to relevant stakeholders and that the impact envisioned is achieved demands for a consortium with a strong record in dissemination, community building and exploitation, and a worldwide network of business partners in various domains.

Table 3.2.1 Key competencies and their coverage and complementarity in QROWD

Participant no.	Participant short name	Linked and open data management	Data quality assessment and repair	Microtask crowdsourcing	Challenges and competition-based crowdsourcing	Collective intelligence and crowdsourcing	GWAPs and gamification	Community initiatives and volunteering-based crowdsourcing	Knowledge modeling, vocabulary engineering and management	Annotation and metadata management	Entity recognition and extraction	Data interlinking and integration	Dissemination, networking and community building	Exploitation	Project management
1	SWC	●●	●		●●			●	●●	●			●●	●●	●
2	BROX	●	●●							●●		●●		●●	
3	ONTOS	●●	●							●	●	●	●●	●●	
4	REEEP	●●	●					●●	●●			●●	●●	●●	●●
5	UNIST	●●	●●		●		●		●●	●●	●●	●●	●	●●	
6	INFAI	●	●●	●	●●	●●	●●	●●	●●		●●	●	●		●●
7	SOTON	●		●●		●●	●●	●	●				●	●	●●

The QROWD consortium brings together seven partners from four European countries (United Kingdom, Germany, Austria and Switzerland), whose profiles, individually as well as in combination, stand for a good coverage of a business focused data value chain and data lifecycle - as it can be seen in the summary in Table 3.3.1. The two research partners are internationally renowned for high quality research and development in Semantic Web, linked data, open data and social computing, and have an impressive track record of successful collaborative research and software development projects, including knowledge and technology transfer and commercialization of scientific results. The industry partners, all of them SMEs, bring in extensive enterprise software development know-how, as well as experience in running industrial projects using LOD technologies (SWC, BROX, ONTOS) and last but not least in the development and successful operation of information and data driven services and products (REEEP, Unister). They will ensure both, that the software components delivered by QROWD are properly anchored in common data practitioner workflows and build vertical showcases demonstrating this in the two business cases that will build on the results of QROWD.

No collaboration with organizations based outside of the EU member states, associated countries, and the list of international cooperation partner countries is foreseen. Subcontracting refers solely to audit costs (see Section 2.4).

3.4 Resources to be committed

The budget plan follows two empirically grounded assumptions: (i) personnel costs have been calculated based on the average person month rates of each partner, excluding overheads; and (ii) the other direct costs are divided into Travel, Equipment, Other goods and services and large research infrastructure. Total project costs, including personnel costs, other costs, and indirect costs, are calculated to 3.630.386.25 € with a requested EC contribution of 2.606.679,88 € (71.80%).

3.4.1 Summary of staff effort

	WP1	WP2	WP3	WP4	WP5	WP6	WP7	Total person/months per participant
1 SWC	5	8	22	30	0	6	18	89
2 BROX	0	12	21	0	20	6	1	60
3 ONTOS	1	10	21	3	7	7	1	50
4 REEEP	0	0	0	32	0	8	1	41
5 UNIST	6	12	0	0	30	6	1	55
6 INFAI	9	22	9	0	8	5	1	54
7 SOTON	20	9	9	5	5	5	1	54
Total person/months	41	73	82	70	70	43	24	

Table 3.3.a Summary of staff effort

The major cost item is by far the personnel costs which accounts for 75.49% of the total budget. Personnel resources are divided into research staff (junior and senior level) as well management staff providing financial and administrative support to the project coordinator.

The management effort is estimated to approximately 0.8 FTE per year throughout the entire duration of the project (6% of total resources). The remaining personnel expenditure is allocated to research and technical staff, who will be involved in development, evaluation, dissemination and exploitation activities.

The effort distribution among partners reflects the very balanced constitution of the consortium and the distribution of roles and activities in the work plan. 72% of the effort is allocated to private sector institutions, broken down as follows: 47,2% for SMEs, 14,2% for a large corporation, and 10,6% to a private NPO. Research institutions hold a share of 28% on the total effort. The project orientation on innovation towards data value chain business is also reflected in the ratio of efforts assigned to technical (WP1,2,3,7) and “near market” work packages (WP4,5,6,7).

Effort distribution across organization types

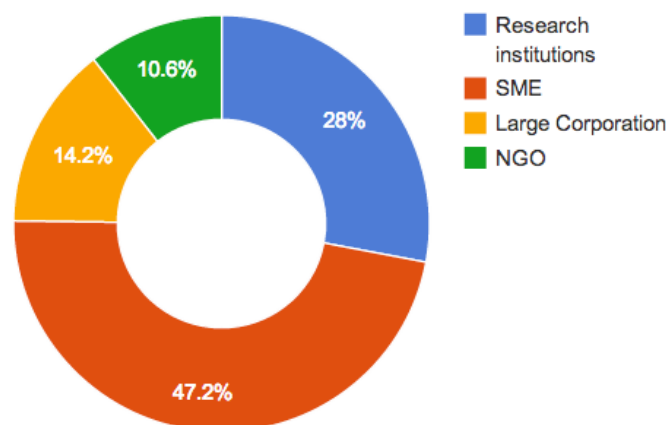


Figure 3.4.1 Effort distribution across organization types

3.4.2 Other direct cost' items (travel, equipment, other goods and services, large research infrastructure)

The other direct costs (4.51% of the total budget) are mostly made up of travel budget for staff to attend project meetings and other events for dissemination and exploitation purposes. We plan

- six plenary meetings with all partners, where
- two meetings of those are co-located with the advisory board meetings,
- three project reviews,
- additional cross-partner meetings where necessary (e.g., work package meetings),
- attendance at international conferences, workshops, and other events.

Server and cloud infrastructure costs are listed under “Equipment” with a total sum for all partners of 23.640,- €. Other goods and services are consumables, audit costs for those partners where the EC contribution exceeds 325K€ with a sum of 40.000 €.

Table 3.4b: Other direct costs for each partner

	1 SWC	2 BROX	3 ONTOS	4 REEEP	5 UNIST	6 INFAI	7 SOTON
Travel	14,748	14,868	16,748	12,948	13,176	13,176	14,580
Equipment	0	0	3,000	0	13,140	7,500	0
Other goods and services	6,500	4,500	3,000	6,500	6,500	8,500	4,500
Large research infrastructure							
Total	21,248	19,368	22,748	19,448	32,816	29,176	19,080

The partners will bring in additional resources and expertise, which will strengthen the project in pursuing its aims and objectives. They will contribute software components and products for the management of different types of content, including unstructured Web documents, structured data (relational as well as graph-based), and linked data (see Section 3.1, WP1 to 3).

3.4.3 Swiss partner

As the negotiations on Switzerland’s association to Horizon 2020 could not be completed, Switzerland has to be considered a non-associated country. For that we treat our Swiss Partner ONTOS as a Third Country party, including all responsibilities and duties with the exception of requesting no EU funding for it (according to Article 9 of the Grant Agreement). Funding for the Swiss partner will be allocated through direct funding by the swiss government. The legal basis for direct funding of Swiss project partners by the Swiss authorities (as was the case before 2004) is mentioned in the Swiss Federal Decree of 10 September 2013 on funding of Swiss participation in European framework programmes in the area of research and innovation for 2014-2020 by the Swiss Parliament, as follows:

Art. 2

Should the financial conditions of the future agreement between Switzerland and the European Union on Swiss participation in EU framework programmes in the area of research and innovation come into effect only after 1 January 2014, then the guarantee credit may be used to cover Swiss participation on a project-per-project basis for the period pending commencement of the agreement.

