Deliverable 8.1.3: Final Release of an Adapted LOD2 Stack for Large Enterprises

Dissemination Level | Public
--- | ---
Due Date of Deliverable | Month 42, 28/02/2014
Actual Submission Date | Month 44, 30/04/2014
Work Package | WP 8, Use Case 2: LOD2 for Enterprise Data Web
Task | T8.3
Type | Prototype
Approval Status | Approved
Version | 0.1
Number of Pages | 20
Filename | D8.1.3.pdf

Abstract: The LOD2 Stack is in its final release. In this deliverable we provide a new vision of enterprise data management by using the latest components of the LOD2 Stack. We show how the LOD2 Stack can be deployed in an enterprise IT system to manage catalogues of linked, semantic and annotated data.

The information in this document reflects only the author’s views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided “as is” without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/her sole risk and liability.
## History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Reason</th>
<th>Revised by</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>27/01/2014</td>
<td>Document creation</td>
<td>Amar Djalil MEZAOUR</td>
</tr>
<tr>
<td>0.1</td>
<td>29/01/2014</td>
<td>Document structure validation</td>
<td>Amar Djalil MEZAOUR</td>
</tr>
<tr>
<td>0.2</td>
<td>04/02/2014</td>
<td>Introduction of deliverable approach</td>
<td>Amar Djalil MEZAOUR</td>
</tr>
<tr>
<td>0.3</td>
<td>23/04/2014</td>
<td>Peer review</td>
<td>Krzysztof Wecel, Sander van der Waal</td>
</tr>
</tbody>
</table>

## Author List

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Name</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXA</td>
<td>Amar Djalil MEZAOUR</td>
<td><a href="mailto:Amardjalil.mezaour@3ds.com">Amardjalil.mezaour@3ds.com</a></td>
</tr>
<tr>
<td>TEN</td>
<td>Bert Van Nuffelen</td>
<td><a href="mailto:Bert.Van.Nuffelen@tenforce.com">Bert.Van.Nuffelen@tenforce.com</a></td>
</tr>
<tr>
<td>OKFN</td>
<td>Sander van der Waal</td>
<td><a href="mailto:sander.vanderwaal@okfn.org">sander.vanderwaal@okfn.org</a></td>
</tr>
<tr>
<td>I2G</td>
<td>Krzysztof Wecel</td>
<td><a href="mailto:krzysztof.wecel@i2g.pl">krzysztof.wecel@i2g.pl</a></td>
</tr>
</tbody>
</table>
Executive Summary

In this deliverable, we describe the next step in our approach in using the LOD2 Stack in an enterprise environment. In fact, in previous deliverables (see D8.3.1 and D8.2.1), we explained how the first release of the LOD2 Stack can be deployed to provide basic semantic features in enterprise data extraction, cleansing, enrichment and interlinking for a dedicated application. In this deliverable we will address another problem area for which enterprises can take benefit of the Linked Data paradigm: the search & maintenance of data sources. Datasets and knowledge about available datasets are the key assets for today’s enterprises success. European governments are stimulating this by actively changing the information landscape. Whereas access to (public) data was in the past mostly paid or simply not available, governments today are opening up data in all areas of social and economic life. This creates new opportunities for enterprises.

This deliverable presents first an overview of the current work in the area of data catalogue management. Next we present the use cases for basic data catalogue management. Then we elaborate more advanced usage scenarios that create data value chains between enterprise and government. Finally we describe the experimental setup that is being constructed to validate the use case scenarios.
## Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD</td>
<td>Linked Open Data</td>
</tr>
<tr>
<td>DCAT</td>
<td>Data Catalog Vocabulary</td>
</tr>
<tr>
<td>EDCAT</td>
<td>Enterprise DCAT API</td>
</tr>
</tbody>
</table>
# Table of Contents

1. **INTRODUCTION** .................................................................................................................. 7

2. **DEALING WITH DATA CATALOGUES: THE BASICS** ................................................................. 8
   2.1 **EXISTING DATA CATALOGUE FRAMEWORKS** .................................................................. 8
   2.2 **LINKED DATA AND DATA CATALOGUES** ......................................................................... 9
       2.2.1 DCAT Vocabulary ......................................................................................................... 9
       2.2.1.1 DCAT-AP vocabulary ............................................................................................... 10
       2.2.2 VoID vocabulary ......................................................................................................... 10
       2.2.3 Registry ...................................................................................................................... 11
       2.2.4 ADMS ........................................................................................................................ 12
   2.3 **INTERCONNECTING DATA CATALOGUES** ......................................................................... 13
   2.4 **RELATED ACTIVITIES** .................................................................................................... 14

3. **LOD2 STACK FOR MANAGING ENTERPRISE DATA CATALOGUES** ......................................... 14
   3.1 **ENTERPRISE DCAT USAGE SCENARIOS** ..................................................................... 15
   3.2 **DISTRIBUTED DATA CATALOGUES** .............................................................................. 16
   3.3 **DATASET ENRICHMENT SCENARIOS** .......................................................................... 16
       3.3.1 Enrichment with annotations with DBpedia concepts .................................................. 17
       3.3.2 Enrichment with annotations with geo-coordinates ..................................................... 17

4. **LOD CATALOGUE ARCHITECTURE** ..................................................................................... 17
   4.1 **LOD2 STACK INTEGRATION IN EDCAT PLATFORM:** ...................................................... 18
   4.2 **GRAPH KNOWLEDGE MODEL** .................................................................................... 18
   4.3 **EVALUATIONS FORESEEN** ............................................................................................ 19

5. **CONCLUSION** .................................................................................................................... 19

6. **REFERENCES** ...................................................................................................................... 20
List of Figures

Figure 1 - Graphical overview of the core concepts in the DCAT standard ............................................. 10
Figure 2 - the core concepts of the registry vocabulary ............................................................................. 12
Figure 3 - The core concepts of the ADMS vocabulary .............................................................................. 13
Figure 3: LOD data catalogue architecture .................................................................................................. 18
Figure 4: Graph knowledge model  ............................................................................................................ 19
Introduction

During the first three years of the LOD2 project, WP8 was fully dedicated to the analysis of the technical possibilities to make linked open data effective in a business application and, more broadly, in an enterprise IT system. Therefore, we collected a set of business requirements to identify the basic expectations of the industry in terms of semantic features, technology needs, type of applications, etc. Based on this, we produced the D8.2.1 deliverable describing generic specifications of what should be an enterprise application implementing linked open data paradigm. To illustrate this deliverable, we implemented the “Active Hiring” application where we demonstrated how linked data can be effective in the specific enterprise application domain of Human Resource management (HR). This implementation was based on the components of the first release of LOD2 Stack.

In this deliverable, the approach of managing enterprise data sets as data catalogues is introduced. The approach consists of a set of use scenarios of how to manage enterprise data catalogues using W3C standard vocabularies, open APIs and formats.

Enterprise businesses manage billions of data items organized in dataset collections and spread over different locations. Some of this data is stored in databases systems, some in file systems and some in specific proprietary systems. From a distance, the collection of these data sources can be viewed as a catalogue or a set of different catalogues when it comes to classifying this data into domains.

This huge amount of data presents for organizations the following challenges:

1. How could we be aware of the existence of data and how to find quickly inside the organization the relevant data?
2. How to exploit this data in novel applications or services?

Since an enterprise is not an isolated entity, but a part of a whole economic eco-system, there are additional challenges:

3. How to find quickly the relevant external data to the organization?
4. How to exploit this data together with the internal data in novel applications and services?

To address these challenges, the enterprise requires a serious amount of effort and the deployment of well-designed and customized search tools. For governmental environments, the open data portals market, like CKAN, OpenDataSoft and Socrata, offers various technical solutions and tools to satisfy the catalogue management requirements. Most of these tools provide advanced features for managing, harvesting and searching data but none of them uses “explicitly” semantic features and W3C standards. The combination of classification/annotation of data and the corresponding faceting facilities at search time offers the impression that the collection of data is browsed in a semantic manner. For example, none of the market tools uses W3C standard formats with explicit semantic links to show similarity relationships for example (with owl:sameAs). The faceting and browsing simulates these relations.

Data catalogues can be the enablers of a “data value supply chain” (challenges 2 and 4). A data value supply chain is the set of enrichments and interlinking processes that happen on a group of datasets. For instance meteorological data is provided by meteorological institutes (many of them are governmental funded), the army, civil air control etc. A farmer weather service could combine all this data into a service that aids farmers in making decisions when to sow, when to water their lands or when to harvest. By making the available data easily discoverable, new data value chains may arise or existing ones get improved (Isbérie & Specty) (Farmer’s Corner).

However, the semantics that is associated with the enterprise data is derived from several elements: machine learning techniques for classification and extraction, domain taxonomies and thesauri, expert validation via editing and authoring tools, etc. These enrichment elements, at a low level (i.e. unit content
by opposition to the overall dataset/catalogue), do not provide a high, general and therefore accurate semantic view and description of the overall collection. It becomes then tricky to query, browse, explore and exploit the collection as it is.

During the previous project years the Linked Open Data context has changed. More tools became available; existing components have been improved; new vocabulary standards have emerged, etc. This created new opportunities. In this last project year we have refocused our integration effort for this use case to explore a new area: namely the exploitation of data catalogues for the enterprise. In our data hungry society, the enterprises that master and exploit best the available information are those that are the most successful.

In our last adaptation of the LOD2 Stack for the Enterprise, we show how the stack components can be used as efficient access, tagging, management tools for enterprise data collections. The key idea of this work is to use (the) standard vocabulary specifically designed for annotating catalogues, DCAT, to group, annotate, enrich, organize and access enterprise data sources. Using LOD2 Stack components create a Linked Data catalogue API for the Enterprise is created, covering above mentioned challenges (1) and (3). The other challenges are more business domain dependent. The ‘Active-Hiring’ prototype application, designed in earlier work of this work package, can be considered as an instance of the two other challenges.

The deliverable is organized as follows:

Section 2 is a survey of the existing work in data catalogue management. It consists of an overview of the state of the art of the existing market solutions. We also describe the relevant vocabularies that are used for data catalogues management.

In section 3 use case scenarios are elaborated for a basic data catalogue management. We provide also extended use cases that deal with enrichment and interlinking of the data catalogues’ content and importantly integration scenarios that explore the integration with external data catalogues. The last scenarios cover challenges 1 and 3 mentioned above.

Finally, in section 4 the solution architecture is presented that will be used to evaluate the use case scenarios.

1. Dealing with Data Catalogues: The Basics

The term Data Catalogue describes a system that manages the descriptions of datasets. A Data Catalogue should be viewed as an electronic library index. It does not organize the data itself, but it structures the descriptions (meta-data) about the actual data.

The term Dataset in the context of a Data Catalogue means the meta data description about the actual data.

1.1 Existing Data Catalogue Frameworks

Data Cataloguing can be achieved with 3 different approaches, by using:

- A dedicated Data Catalogue
- A Web Portal
- A Document/Content Management systems

Dedicated Data Catalogues are very popular with governments and public organizations which allow organizations to disseminate efficiently the availability of (their) data. In this category, CKAN is the most prominent. It is an open-source community project providing an extensible Data Catalogue framework
[REF]. Other open-source alternative is DKAN, a Drupal extension (Drupal DKAN), while commercial solutions are (OpenDataSoft) and Socrata (Socrata).

Although these frameworks cover the majority of so called Open Data Portals, most of existing data portals are classical web portals. The latter are often not recognizable as data portals because they often exploit the notion of a dataset implicit in their portal design and focus on the story telling around the data while hiding the data in a hard to reach corner. Open data portals do not focus on the storytelling, but merely on the dissemination of the availability of the data.

The third approach is the application of a document management system. Within enterprises document management & collaboration frameworks are commonly setup. Although not designed with being a data catalogue in mind, they can be turned in a data catalogue. A prototypical example is Microsoft SharePoint, where documents are tagged with ownership, version and user added information. Via the search functionality users of the internal Microsoft SharePoint can find easily the needed documents.

Although each approach focusses on an aspect of dataset sharing, they include elements of the others.

1.2 Linked Data and Data Catalogues

Open Data Portals, hence Open Data, and Linked Data are often mentioned as part of the same story. Although technically they are not connected, both tackle similar problems at different levels. On the one hand, Open Data Portals aim at easing the access to Data. In that support the access is typically underpinned with legal support for the use of the data by Open Data licenses\(^1\).

Linked Data, on the other hand, is technically opening data by a general purpose data representation format RDF. RDF is intrinsically opening up the data because it is application neutral. Information encoded in RDF is ready to be processed by any RDF-aware application. It is also storage neutral, moving the storage from an in-memory store to OpenLink Virtuoso\(^2\) and then to Ontotext OWLIM store\(^3\) is as simple as dumping the data in text files and uploading them again in the target store.

Being a general purpose data carrier is not sufficient to be an enabler for information sharing and application reuse. This is enabled by the semantic layer associated with RDF. By giving meaning to classes and properties and sharing this meaning, data reuse is enabled. This is the crucial role that vocabulary standards play. For Data Catalogues the key W3C standards are discussed below.

To our knowledge, CKAN is the only data catalogue system that provides some support for offering the data catalogue content as Linked Data. CKAN supports the export of the dataset description in RDF. See http://docs.ckan.org/en/latest/linked-data-and-rdf.html. Native Linked Data support covering dereferenceable URI’s and a SPARQL-endpoint are not present. This can be created by setting up an RDF store that is synchronized with the CKAN instance. An example of such setup is the European Open Data Portal. (http://open-data.europa.eu).

1.2.1 DCAT Vocabulary\(^4\)

DCAT is a W3C recommendation. This is the base vocabulary enabling the interoperability between data catalogues. Quoted from the standard:

"DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. [...]"
By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites.

Figure 1 - Graphical overview of the core concepts in the DCAT standard

1.2.1.1 DCAT-AP vocabulary

The DCAT-AP profile is a specialization of the DCAT standard. It is aimed towards governmental Open Data portals. Typical specializations are the usage of some controlled vocabularies as the range for some properties. For instance “dcat:theme”’s range has been fixed to the 20 top concepts from the Eurovoc thesaurus (http://eurovoc.europa.eu/).

1.2.2 VoID Vocabulary

This vocabulary is aimed at describing RDF datasets.

Quoted from the abstract:

---


6http://www.w3.org/TR/void/
“VoID is an RDF Schema vocabulary for expressing metadata about RDF datasets. It is intended as a bridge between the publishers and users of RDF data, with applications ranging from data discovery to cataloguing and archiving of datasets. This document is a detailed guide to the VoID vocabulary. It describes how VoID can be used to express general metadata based on Dublin Core, access metadata, structural metadata, and links between datasets. It also provides deployment advice and discusses the discovery of VoID descriptions.”

This vocabulary is more restrictive than DCAT because of its focus on only RDF datasets. But at the same time, this focus allows to express more properties of interest. Important aspects being supported by VoID are expressing the vocabulary used in the dataset and the ability to express dataset partitions. In addition VoID offers support for expressing information for linksets. A linkset is a dataset containing links between two datasets. Within the Linked Data community a linkset is a very common and important kind of dataset as it embodies the whole idea of linking data with each other.

1.2.3 Registry

This vocabulary is closely related to a data catalogue. Although the terminology is different, the structural approach is very similar. One of the first example usages is about a register containing VoID dataset descriptions. The vocabulary is, however, intended for a wider usage than DCAT. A register collects not only datasets, but it collects any kind of information unit. The work around versioning and register federation is of interest, two aspects that are not so deeply elaborated in DCAT.

---

7 http://purl.org/linked-data/registry
8 See https://github.com/UKGovLD/ukl-registry-poc/wiki/Extensions#dataset-registration
1.2.4 ADMS

Quoted from the introduction

ADMS, the Asset Description Metadata Schema, is a profile of DCAT for describing so-called Semantic Assets (or just ‘Assets’), that is, highly reusable metadata (e.g. xml schemata, generic data models) and reference data (e.g. code lists, taxonomies, dictionaries, vocabularies) that are used for eGovernment system development. Someone searching for an Asset is likely to have different needs, priorities and expectations than someone searching for a dataset in a data catalogue and these differences are reflected in ADMS. In particular, users seeking an Asset are likely to be searching for a document — something they can open and read using familiar desktop software, as opposed to something that needs to be processed. Of course this is a very broad generalization. If a code list is published as a SKOS Concept scheme then it is both an Asset and a dataset and it can be argued that all Assets are datasets. Therefore the difference in user expectation is at the heart of what distinguishes ADMS as a profile of DCAT. A further distinction between DCAT and ADMS can be made in that DCAT is designed to facilitate interoperability between data catalogues, i.e. the catalogue itself is at the heart of the vocabulary. ADMS is focused on the assets within a catalogue.

---

9 http://www.w3.org/TR/vocab-adms/
1.3 Interconnecting Data Catalogues

In today’s world organizations are not isolated monolithic entities. Although large organizations may be perceived externally as a single unit, e.g. the European Union or Dassault Systèmes, they are in fact an agglomeration of individual units. Data catalogues should therefore, instead of enforcing the existing organization’s boundaries, respect the distributed nature of the organization while at the same time make sure that the information is reachable for all interested parties.

(Base) data catalogues are commonly setup as closed units that expose information received for a group of dataset publishers. For instance, the European Open Data Portal contains the information from all European institutes; the British Open Data Portal (data.gov.uk) contains the information from the United Kingdom. Other portals are thematic and contain for instance only meteorological data.

Although such portals ease the search for a dataset inside its covered domain areas, the hunt for a dataset of interest has now shifted towards which portals to interrogate. Typically the search for data is not limited to one single portal. This problem is tackled by the creation of meta Data Catalogues. These meta Data Catalogues harvest the content of other data catalogues. They are not the prime owner of the dataset descriptions, but construct a view on them. CKAN examples are publicdata.eu and datahub.io, a Linked Data (DCAT-AP) data catalogue example is: http://data.opendatasupport.eu. To our knowledge, there are no approaches present that support the distributed nature of data catalogues.
1.4 Related Activities

After having introduced the existing solutions, the core vocabularies and the approaches used for in interconnecting data catalogues, we conclude the overview of data catalogues with related work that has happened outside and within the LOD2 project.

Within the LOD2 project, several initiatives and elaboration tracks have been set up and can be seen as closely related to this work. Foremost WP9 is closely related as its main objective is to create and stimulate a more open and accessible space for public data. PublicData.eu has been created to highlight all Europa’s public data. This data catalogue is a CKAN instance which harvests over 30 public governmental data portals (mainly other CKAN instances).

Viewed from a much wider perspective, the European Union is actively stimulating open data in all its aspects. This has as effect that all over Europe governments take action at two levels:

- Firstly, they are providing more and more access to data free of charge or at marginal costs, see (Commission, Implementation of the Public Sector Information Directive) and (Support).
- Secondly, they are creating Open Data Portals, see (Commission, European Union Open Data Portal) which increase the visibility of the available data.

For enterprises this evolution creates opportunities. In order to further stimulate this area, funding is provided via research and innovation platforms like FP7 and Horizon2020. For reference, some pointers to past and ongoing related projects COMSODE\(^{10}\) (COMSODE), ENGAGE\(^{11}\) (ENGAGE), smart cities\(^{12}\) (SmartCities), Open Data monitor\(^{13}\) (OpenDataMonitor), open data support\(^{14}\) (Commission, Open Data Support).

We have not encountered any projects that focus on creating data catalogues targeted on enterprises. This is indirectly supported by current focus of the Horizon 2020 related calls due in 2014 for funding on the creation of a data value chain. An initiative worth mentioning in LOD2 related to this case is the integration application for European Legal data. It interconnects the European Publication Office SPARQL endpoint and the Wolters Kluwer knowledge base setup in WP7. Internally in the integration application a DCAT catalogue is setup over the provided data. This harmonization has enabled the creation of more universal applicable data selectors; hence simplification of internals of the application (see Deliverable D7.2.2). One of the recommendations from this exercise is to consider DCAT as a dissemination vocabulary for the Publications Office SPARQL endpoint content.

2. LOD2 Stack for Managing Enterprise Data Catalogues

In the following section, we describe an LOD based data catalogue solution. To introduce this solution, we elaborate, first, the justification to create a native LOD data catalogue instead of extending an existing system such as CKAN.

**Our objectives:**

\(^{10}\) [http://www.comsode.eu/](http://www.comsode.eu/)
\(^{11}\) [http://www.engage-project.eu/wp/](http://www.engage-project.eu/wp/)
\(^{12}\) [http://eu-smartcities.eu/](http://eu-smartcities.eu/)
\(^{13}\) [http://project.opendatamonitor.eu/](http://project.opendatamonitor.eu/)
\(^{14}\) [http://opendatasupport.eu](http://opendatasupport.eu)
**DCAT compliance**: Our main objective is to have a Linked Data catalogue completely in line with the DCAT standard. DCAT is selected as from all presented vocabularies it has the largest market penetration.

This objective represents for existing data catalogues like CKAN a serious effort on many aspects of the framework. For instance according to the DCAT standard a contact point is a “foaf:Agent”, which is therefore a reusable and unique addressable item. In CKAN a contact point is not a unique item, but a collection of fields associated with a dataset. For such situations and other reasons, Open Knowledge Foundation has not placed a native DCAT compliant objective on the CKAN roadmap, but provided a mapping from the CKAN content to DCAT for each dataset.

**Interoperability using linked data formats**: For enterprises interoperability between different systems is crucial. Having different components, usually developed by different vendors, of an ERP or CRM or any enterprise application talking to each other is important for ensuring the efficiency of information processing. This objective is nowadays done at the technological levels where software wrappers and intermediate communication buses are used to enable interoperability. This can be achieved more efficiently if the applications use standard formats, specifically RDF W3C standards.

**Data linking and enrichment functionalities exploitation**: The LOD2 Stack is a collection of tools which share the same base: namely the data format on which they act is RDF. A RDF data catalogue will be able to exploit all these functionalities. Notable LOD2 Stack properties that are relevant for the enterprise are the distributed nature of the data storage, the enormous scaling potential with Virtuoso7 and the uniform management of non-catalogue the enrichment and interlinking data in coherence with the catalogue data.

**Enterprise system integration**: Integration of software inside an enterprise is typically driven along two dimensions: the system level dimension and the technology dimension. The system level dimension is roughly split in two categories. In the first case the functionality is best offered in the UI context already existing at the customer side; in the latter case, the functionality is integrated in the existing data processing pipelines. For the UI integration theming, the MVC-paradigm and an efficient API to fetch individual pieces of information are required; for data processing pipelines efficient storage layer access with a focus on storage and validation functionalities is required.

The technology dimension is driven by the need of the enterprise to maintain the integrated data catalogue. Many enterprises have decided to work around a technology kernel for reducing knowledge drain risks. Since RDF and SPARQL are not the most common technologies, we will provide a JSON interface that opens up the data catalogue to software engineers having knowledge of DCAT as data model. The (first version of the) API will cover CRUD operations\textsuperscript{15} on datasets and structural and free text search.

**Supporting the data value chain**: Although the interconnection of data catalogues seems to be an obvious consequence offering a large additional value for a catalogue user, there is little knowhow and work done around this topic. All known approaches (publicdata.eu and data.opendatasupport.eu) exploit harvesting. Harvesting is the process which copies and transforms other catalogues content at regular times. (SPARQL) federated queries are here a valid alternative. With this technique queries are evaluated over a number of data catalogues. This eliminates the need for copying data, but requires a more intelligent query answering system and imposes a higher level of standardization on the data catalogues.

### 2.1 Enterprise DCAT Usage Scenarios

Our objectives are translated in the following use case scenarios:

1. Elementary operations on data collections to group them into logical units (catalogues and datasets)

\textsuperscript{15} \url{http://en.wikipedia.org/wiki/Create,_read,_update_and_delete}
2. Labelling and semantic annotation of datasets/catalogues
3. Datasets and catalogues can be interlinked
4. Datasets can be manipulated via a JSON REST interface. This JSON REST interface output is DCAT vocabulary compliant.
5. Datasets should be searchable via free text search functionality.
6. Data catalogues should be able to add easily proprietary information to the JSON REST interface.
7. Information about datasets can be retrieved in multiple formats.
8. Before data is ingested in the data catalogue the information is validated.
9. Validation rules can be managed without changing the software
10. Access URLs are checked for existence
11. Data quality is measured
12. The framework can handle multiple data catalogues in one store.
13. Support for importing existing DCAT based data catalogues.
14. System administrators can track the API usage and execution by inspecting a log.
15. The API is easily extendable to incorporate data catalogue specific assumptions.

These 15 use cases form the cornerstone of our requirements for a LOD data catalogue based on the DCAT vocabulary. The last 3 use cases are necessary steps for being able to deploy such API in an enterprise.

2.2 Distributed Data Catalogues

Although the creation of a data catalogue holds value for the enterprise, the value of a data catalogue is multiplied with the creation of more internal and external data catalogues. RDF is as data format well-suited for distributed data; the federated queries extension to SPARQL\[^{16}\] is enabling this potential at the query level.

The scenarios that will be explored with this extension are:

1. A data catalogue within an enterprise and a governmental data catalogue
2. Scenario 1) extended with the addition of an extra governmental data catalogue
3. Scenario 1) extended with splitting of the enterprise data catalogue in two
4. Scenario 3) extended with splitting the public data catalogue in two
5. Starting from Scenario 4) remove one by one the external data catalogues.

Scenario 1 covers the typical integration case: an external company is exploiting governmental data catalogues for their business. The setup applies also for larger companies where instead of an external government data catalogue a data catalogue from another part of the company is included. Therefore, intranet and extranet situations are covered with the same technology.

Data catalogues are a living landscape. Therefore typical network changes must be investigated. Scenario 2, 3, 4 and 5 cover four such changes.

2.3 Dataset Enrichment Scenarios

Having data catalogues in linked data is an enabler to exploit the potential of the LOD2 Stack. Whereas the above scenarios are concerned with data collection, data quality and data access, in this section we will describe scenarios that will exploit the information in a more innovative way.

\[^{16}\] \url{http://www.w3.org/TR/sparql11-federated-query/}
2.3.1 Enrichment with Annotations with DBpedia Concepts

Based on the dataset’s title and descriptions, external concepts such as those from DBpedia can be associated with the dataset. This enrichment creates a potential to explore the data catalogue with an alternative classification.

2.3.2 Enrichment with Annotations with Geo-Coordinates

Geospatial information is often available in the dataset metadata. For instance, the location of the publisher, the geospatial coverage of the data catalog dataset and dataset distributions. Based on this information geo-coordinates can be associated and these can be used to explore the data catalogue via a geospatial search.

Since such enrichments form a corner stone in Linked Data, the Enterprise DCAT API (EDCAT) platform is to be expected to offer support with such enrichments.

3. LOD Catalogue Architecture

The use-cases described in the previous section are covered by the following logical architecture of our data catalogue. It is depicted below. The basis is the storage layer. The master storage is by an RDF store, which is domain agnostic. Free text search will be offered via a search index which content is regularly synchronized with the content of the RDF store. Also here a domain agnostic solution is chosen. Distributions can be stored on server if necessary via a WEBDAV.

The Enterprise DCAT API (EDCAT) is the interface to the data catalogue. This layer embeds minimal domain knowledge. A first way of access is given via the provided SPARQL endpoint and the dereferenceable URI’s. Next access is via content negotiation: not only meta data can be retrieved via content negotiation, but the data itself can be retrieved. The last kind of access is the JSON API. This API supports CRUD and free text search operations.

The API is supported by system management and business management functionalities:

1. Business level:
   1. Data validation: data inserted in the store must satisfy the mandatory constraints and any optional constraints. The optional constraints aid to detect quality related issues: with this information publishers are able to make sure their datasets are more easily found.
   2. Data quality assessment: In addition to the data validation service, the value of the data catalogue is improved if the quality of the available data is monitored.

2. System level:
   1. Authentication & authorization: controlling the accesses to the data catalogue (read and updates) is important
   2. Logging: in order to understand the API interaction with the storage layer logging is important. This allows identifying problems like service bottlenecks.

---

17 http://en.wikipedia.org/wiki/Webdav
3.1 LOD2 Stack Integration in EDCAT Platform:

The DCAT Enterprise platform consists of the following LOD2 components:

1. RDF Store: Virtuoso
2. WebDav file store: Virtuoso
3. Search Index: SIREn
4. Enterprise DCAT API: new software code: Java, OpenRDF Sesame, SPARQL

3.2 Graph Knowledge Model

An important aspect of any Linked Data application is the understanding of the applied graph structure. The EDCAT API configuration is located in its configuration graph. If the API requires configuration for a specific functionality then this information is stored in an independent graph. For instance, the validation service is a collection of Validation Rules with information about which rules are active for a catalogue. The EDCAT is capable to handle multiple catalogues. A catalogue is an independent graph containing general catalogue information and the catalogue records of the contained datasets. The datasets themselves are also stored in an independent graph, hence a graph per dataset.

Although this last design choice leads to an increasing number of graphs, the benefit is that updates to datasets are well-scoped. Implicit data sharing that would happen when storing in one graph will not occur. A typical example is the reference to a contact point. Deletion of a dataset corresponds to the deletion of a dataset graph; hence the reference to (and possible details of) a contact point disappear. When the information was stored in a single graph, the deletion operation would have to figure out if the contact point details are solely associated with the dataset or shared with other datasets. In the last case the removal of contact point details cannot be done.
3.3 Evaluations Foreseen

With this application we will be able to evaluate some LOD2 Stack components in a different context than what has been done in other use cases. Especially the evaluation of SIREN is of interest as this is the first LOD2 Stack integration in which SIREN is applied. Testing data will be taken from the Open Data Support project catalogue (data.opendatasupport.eu).

Since this work is oriented towards businesses, it is our aim to let our work to be evaluated by some potential enterprise users.

4. Conclusion

Data catalogues are the information crossroads on Europe’s data highway. In the past years the European Union has invested a lot in the opening of data and it is laying down the foundations of a
European data highway network. The legal framework of this effort is covered by initiatives such as PSI and Inspire. Technical interoperability is boosted by funding CONSODE and LOD2. So far, these initiatives concern mostly governmental organizations and much less private businesses.

Our work described in this deliverable covers exactly this second aspect: how enterprises can exploit and join the data highway that Europe is creating. We have described the current approaches to increase the interoperability between the data catalogues. Based on the DCAT vocabulary, we have outlined the needs for a REST API covering that area. A solution architecture based on the LOD2 Stack has been presented. This solution will be evaluated in the next months and the outcome will be reported in the final evaluation deliverable D8.3.2.

5. References

COMSODE. Components Supporting the Open Data Exploitation. http://www.comsode.eu/
Drupal DKA. https://drupal.org/project/dkan
Isbérie, C., & Specty, R. How farmers use information obtained from irrigation advisory services in France. http://www.fao.org/docrep/w4367e/w4367e07.htm