5.2 CODE Platform Beta Version

This document summarizes the CODE Platform that has been build. The platform consists of three prototypes to analyse different value chains for (Linked) Open Data:

- [42-data.eu](http://42-data.eu) – a data-centric question and answer portal for people believing in data
- The Mendeley Desktop/Client and Server API for semantically enriching research publications.
- MindMeister.com web platform for generating semantic web enabled mind-maps and mind-map-based presentations

All three platforms are characterized by their production and consumption of Linked Open Data. They target to evaluate the value of linked open data within research and data intensive processes.

<table>
<thead>
<tr>
<th>Project Acronym</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grant Agreement number</td>
<td>296150</td>
</tr>
<tr>
<td>Project Title</td>
<td>Commercially Empowered Linked Open Data Ecosystems in Research</td>
</tr>
<tr>
<td>Date</td>
<td>2014-02-05</td>
</tr>
<tr>
<td>Nature</td>
<td>P (Prototype)</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU (Public)</td>
</tr>
<tr>
<td>WP Lead Partner</td>
<td>Mendeley</td>
</tr>
<tr>
<td>Revision</td>
<td>9th version - final</td>
</tr>
<tr>
<td>Authors</td>
<td>Florian Stegmaier, Michael Granitzer, Patrick Höfler, Roman Kern, Kris Jack, Michael Hollauf, Vedran Sabol</td>
</tr>
</tbody>
</table>

Consortium:

Statement of originality: This document contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. This document reflects only the author’s views and the European Community is not liable for any use that might be made of the information contained herein. © CODE Consortium, 2012
Project Officer & Project Coordinators

Project Officer | Stefano Bertolo | European Commission  
                  | DG CONNECT Unit G3  
                  | EUROFORUM 01/293 - 10 rue Robert Stumper, L-2557 Luxembourg  
                  | stefano.bertolo@ec.europa.eu

Project Coordinator | Stefanie Lindstaedt | Know-Center GmbH  
                    | Inffeldgasse 13, 8010 Graz, Austria  
                    | +43 316 873-30800 (phone)  
                    | +43 316 873-1030800 (fax)  
                    | slind@know-center.at

Scientific Coordinator | Michael Granitzer | University of Passau  
                        | Innstrasse 33a, D-94032 Passau  
                        | +49(0)851-509-3305  
                        | michael.granitzer@uni-passau.de

Document Revision History

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Author</th>
<th>Organization</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st} draft</td>
<td>26.01.2014</td>
<td>Florian Stegmaier, Michael Granitzer</td>
<td>University of Passau</td>
<td>Initial Q&amp;A platform description</td>
</tr>
<tr>
<td>2\textsuperscript{nd} draft</td>
<td>28.01.2014</td>
<td>Patrick Höfler</td>
<td>Know-Center</td>
<td>Query Wizard and Vis Wizard integration into the Q&amp;A platform</td>
</tr>
<tr>
<td>3\textsuperscript{rd} draft</td>
<td>03.02.2014</td>
<td>Florian Stegmaier, Michael Granitzer</td>
<td>University of Passau</td>
<td>Q&amp;A platform description completed, success indicators, introduction</td>
</tr>
<tr>
<td>4\textsuperscript{th} draft</td>
<td>04.02.2014</td>
<td>Roman Kern</td>
<td>Know-Center</td>
<td>CODE Annotator Tool</td>
</tr>
<tr>
<td>5\textsuperscript{th} Draft</td>
<td>04.02.2014</td>
<td>Michael Granitzer</td>
<td>University of Passau</td>
<td>Introduction</td>
</tr>
<tr>
<td>6\textsuperscript{th} Draft</td>
<td>05.02.2014</td>
<td>Kris Jack</td>
<td>Mendeley</td>
<td>Mendeley Desktop</td>
</tr>
<tr>
<td>7\textsuperscript{th} Draft</td>
<td>05.02.2014</td>
<td>Michel Hollauf</td>
<td>MindMeister</td>
<td>Semantic Mind Maps</td>
</tr>
<tr>
<td>8\textsuperscript{th} Draft</td>
<td>05.02.2014</td>
<td>Vedran Sabol</td>
<td>Know-Center</td>
<td>Integration, layout, proof reading</td>
</tr>
</tbody>
</table>
9th - Final version | 05.02.2014
Table of Contents

1 Introduction - 4 -

2 Mendeley Desktop and Server API - Enhanced Data Extraction from PDFs - 5 -
  2.1 TABLES OF CONTENTS - 6 -
  2.2 TABLES - 7 -
  2.3 FIGURES - 8 -
  2.4 ACCESS VIA APIs - 8 -
  2.5 CODE ANNOTATOR TOOL - 9 -

3 Semantic Mind Maps - 13 -
  3.1 WUNDERKIND: BOOSTING MIND MAPS THROUGH LINKED DATA - 13 -
  3.2 PUBLISHING SEMANTIC MIND MAPS IN THE LOD - 17 -
  3.3 FROM MIND MAPS TO PRESENTATIONS - 20 -

4 42-data: A data centric question and answering portal - 22 -
  4.1 42-DATA: DATA IN CONTEXT - 22 -
  4.2 THE DONATION AND REPUTATION CONCEPT - 24 -
  4.3 AVAILABLE RESOURCES IN 42-DATA - 24 -
  4.4 INTEGRATED WORKFLOWS IN 42-DATA - 28 -
  4.4.1 CREATION OF DATA CUBES ON THE BASIS OF (LINKED) OPEN DATA - 28 -
  4.4.2 CREATING QUESTIONS AND ANSWERS — A RESOURCE CENTRIC APPROACH - 30 -
  4.4.3 BOOKMARKING RESOURCES FOR CREATING YOUR OWN DATA LIBRARY - 32 -

5 Success Factors - 33 -

6 References - 35 -
1 Introduction

CODEs vision is to establish the foundation for a web-based commercially oriented ecosystem around Linked Open Data. We focused on research data as an underlying application scenario.

In order to realise this vision, we worked on two kinds of ecosystems:

1. Linked Data-enhanced services to empower existing platforms
2. A linked data platform for allowing Linked-Data laymen to aggregate, analyse, share and collaborate on Linked Open Data.

This deliverable describes the prototypes/platforms for the above ecosystems and shows their usage in a Handbook like manner.

The first platform is the Mendeley Desktop Client and the corresponding Mendeley API. The Mendeley Desktop Client uses CODE services that enrich scientific research papers by extracting table of contents, tables, figures and entities for nearly one million open access publications available in Mendeley. The extracted elements are partially available as JSON-LD and freely available via the Mendeley REST API. The features are available to all Mendeley Users and it is the first time that enriched PDFs are available to the scientific community.

The second platform is the MindMeister Collaborative mind-mapping platform. Through CODE MindMeister now provides a semi-automatic extension of MindMaps with different Linked Open Data Sources like Freebase. Mind-maps, which constitute thesauri, are exported using the SKOS Vocabular and hence are available for further re-use in the semantic web. For research papers, MindMeister offers a presentation mode on top of enriched research papers and CODEs linked data search and visualisation facilities.

The third platform, which we named 42-Data, integrates all CODE services in order to bring Linked Open Data to Linked Data laymen. 42-Data constitutes a data bookmarking and data-centric question and answering portal. Researchers are enabled to explore the Linked Open Data cloud and to collect data from different sources like research papers, or non-semantic open data portals. With the collected data they can engage in discussions using data to underpin their opinion. The portal emphasizes a donation based revenue model. The model has been chosen to allow micropayments among participants without enforcing legal contracts for requesting/delivering data.

Finally, we also report on the currently achieved success factors. However, since the launch of the platforms have been delayed, we did not achieve all success factors yet.
2 Mendeley Desktop and Server API - Enhanced Data Extraction from PDFs

Mendeley’s mission is to open up research in order to help researchers better organise their research, collaborate with one another and discover new research. Mendeley produces tools that help researchers with these tasks. Many of these tools consume and produce data that is stored on Mendeley’s platform. As Mendeley’s community grows and more people make use of the tools offered, the richer the data becomes. It also powers new tools, such as recommender systems, that require such rich data collections. Mendeley is opening up much of this data through their API so that third parties can also build tools that help researchers using the data that has been collected and perhaps also feeding back more data.

Before the CODE project began, Mendeley had not used linked data. The CODE project allowed Mendeley to investigate how linked data could be used in order to further open up research. It proposed to link up the contents of research articles to the Open Linked Data Cloud (OLDC). This could be done by connecting entities that appear in the text, table of contents, tables and figures of articles with the OLDC. Part of this process would be through automatically making these links while the other part of the process would be crowdsourcing it by allowing Mendeley users to make the links. This would allow researchers to not only consume linked data but also to contribute back to it. Finally, these services would be made available on a common platform, such as Mendeley’s API, in order to also open linked data up to developers.

Thus far in the project, Mendeley has integrated a number of the CODE project services into its main product, Mendeley Desktop, and its API. Mendeley Desktop has so far been downloaded by over 2.5 million people throughout the world and is regularly used by many of them. The services integrated into Mendeley Desktop are provided for free to Mendeley users and are exposed through Mendeley’s API, supporting the platform model. Where services developed through the CODE project are available, a link is included to a user-friendly description of them and accreditation is given to the CODE project, its participants and funders.

This chapter includes a description of the CODE services that have been integrated into Mendeley Desktop and Mendeley’s API. These services automatically extract the following data from pdfs:

1. Tables of Contents
2. Tables
3. Figures
4. Entities

Instructions are provided on how to use the services, which, as of 1st of February, 2014, are publicly available as development previews and will be pushed as an automatic update to all Mendeley Desktop clients before the end of February, 2014. Around one million Open Access articles have been processed. For ease of finding Open Access articles, Mendeley Desktop provides a search filter that can be manually selected from the ‘Literature Search’ option (Figure 1). Instructions for using these tools now follow.
2.1 Tables of Contents

![Example of a table of contents extracted from a pdf.](image)

**Figure 2: Example of a table of contents extracted from a pdf.**
Mendeley Desktop shows the table of contents that have been extracted by KNOW-Center’s extraction service. The headings in the table of contents become links that you can click on to take you directly to those sections in the pdf viewer.

1. Open up a pdf in Mendeley Desktop’s internal pdf viewer. In the right hand, there are four tabs: “Details”, “Notes”, “Contents” and “Summary”.
2. Click on the “Contents” tab. Mendeley Desktop will show the table of contents from the pdf in terms of the headings and subheading contained within it (see Figure 2).
3. Click on one of the headings that has been extracted to jump directly to that point in the article.

2.2 Tables

![Figure 3: Examples of tables extracted from a pdf.](image)
Mendeley Desktop shows the tables that have been extracted out by the KNOW Center’s extraction service. The table becomes a link that you can click on to take you directly to it in the pdf viewer. Not only is the table extracted but its contents are also recognised, allowing them to be copied and reused in other applications.

1. Open up a pdf in Mendeley Desktop’s internal pdf viewer. In the right hand, there are four tabs: “Details”, “Notes”, “Contents” and “Summary”.
2. Click on the “Summary” tab. Mendeley Desktop will show the tables that are contained within it (see Figure 3).
3. Click on one of the tables that has been extracted to jump directly to it in the article.
4. Click on the arrow to the top right of a table in order to reveal the option to copy its contents.

2.3 Figures
Mendeley Desktop shows the figures that have been extracted out by the KNOW Center’s extraction service. The figure becomes a link that you can click on to take you directly to it in the pdf viewer.

1. Open up a pdf in Mendeley Desktop’s internal pdf viewer. In the right hand, there are four tabs: “Details”, “Notes”, “Contents” and “Summary”.
2. Click on the “Summary” tab. Mendeley Desktop will show the figures that are contained within it (see Figure 4).
3. Click on one of the figures that has been extracted to jump directly to it in the article.

Figure 4: Example of figure extracted from a pdf.

2.4 Access via APIs
The tables of contents, tables and figures that have been extracted using the KNOW-Center’s services are available through Mendeley’s API (toc endpoint in Figure 5). The entities that were extracted
using the KNOW-Center’s services and disambiguated using Passau’s services are also available (entities endpoint in Figure 5). For general instructions on how to use the API, please consult http://dev.mendeley.com/alpha-release/.

Figure 5: Example of the endpoints currently available in Mendeley’s API

2.5 CODE Annotator Tool

In order to demonstrate the algorithmic underpinning of the work conducted for enriching PDFs [Kern2013] we developed a demonstration application. This application allows us to quickly integrate workflows and methods to extract facts from PDFs. The design of the tool targets “power-users”, i.e. experts with understanding of how extraction algorithms work and which requirements must be fulfilled for them to operate. In the following Figure 6 to Figure 11 demonstrate how the CODE Annotator Tool is used.

Figure 6: Initial CODE Annotator Tool Web-page (left) and the user log-in (right).
Data Sets
To start your annotations you need to first select a data set. If no suitable data set is installed, you can simply import one into your account.

Add data sets
You may choose to import new data sets to your account. Once a data set is imported it can be selected from the list of installed data sets.

Select from installed data sets
Select one of the installed data set

Figure 7: Annotation process begins by selecting an existing data set or importing a new one, for example from Mendeley.

Figure 8: A model can be created by importing a taxonomy specified in MindMeister. Alternatively an existing model can be used.
Figure 9: Expert view: tuning the parameters of the model. Subsequently the model can be applied to annotate other data sets or to evaluate its accuracy.

Figure 10: Find the document to be annotated using the semantic search.
Figure 11: Annotate the document for the selected model.
3 Semantic Mind Maps

MindMeister is the leading web-based mind mapping tool and has currently close to 2.5 million registered users. The company’s vision has always been to bring mind mapping – which has always been, and for all intents and purposes is a niche concept, to the masses. We want to achieve this by focusing on a very simple and easy-to-use end user interface of our products, by removing unnecessary and confusing functionality and concentrating on the essentials, and by using modern technologies to provide as much mapping help to our users as possible.

The first two goals have been achieved in the 7 year history of the company, while the third goal – to use technology to facilitate better and fast mind mapping – we’re still struggling with. We’ve introduced concepts such as WunderLink (auto-link topics to web pages), WunderNote (auto-retrieve Wikipedia descriptions for topics) and WunderImage (auto-enhance maps with matching graphics). The larger goal, however, was to provide relevant, contextual content suggestions to our users while mapping, using structured resources from the web. We joined the CODE project in the hope that linked data could provide the missing piece of the puzzle.

Furthermore, mind maps are inherently a visualization of linked data. The tree-like structure and the ability to draw connections between nodes creates semantic relationships between the topics. To visualize this information, and to provide a way for feeding it back into the linked data cloud, was another goal of our participation.

Finally, the MindMeister presentation mode provides an attractive way of visualizing interlinked information as effect-driven slideshow and differentiates itself from classic PowerPoint-style presentations by maintaining structure and context for the user during the course of the presentation. We aimed at providing easy ways for researchers to turn papers into online presentations, using automated extraction tools and enrichment with LOD visualizations.

At this stage in the project we’re proud to say that we’ve successfully implemented modules for all three scenarios and included them in the beta version of the CODE platform. These are:

- Wunderkind: provide contextual suggestions for child nodes in mind maps
- RDF Export: publish semantic mind maps to the LOD cloud
- PDF to Presentation Wizard: create online slideshows from research papers

In the current stage, these features will be made available to our users via an “Experimental Features” setting that all users can enable. We do this to gather insights about the usage of the features as well as their quality and relevance for our user base. Depending on these outcomes they will be incorporated into the main feature set.

MindMeister uses the Freemium business model, i.e. we provide our basic services free of charge but provide advanced features and unlimited map creation to paying users. As such it is important to keep developing cutting-edge functionality to entice users to upgrade to a paid plan. We think that the features developed for the CODE beta platform should be paid features, provided they work as expected in practice.

The remainder of this chapter describes the developed modules in detail.

3.1 Wunderkind: Boosting Mind Maps through Linked Data

As explained above, the idea behind Wunderkind was to allow for automated, contextual content suggestion by providing MindMeister users a virtual partner to collaborate with while building mind maps. In the first version, the virtual partner would provide suggestions only when triggered by user
interaction, although this can be expanded in the future to a more intelligent behaviour of showing relevant data as the user is adding new content to the mind map in a “Experimental Features” mind map mode.

![Diagram](image)

**Figure 12: Current flow of the Wunderkind interaction**

The figure above shows the data flow of the Wunderkind feature. At most, there are three calls made to the respective endpoint that provides the suggestions based on the selected topic of the mind map. A typical request for Wunderkind is detailed below.
Figure 13: The user opens the suggestions panel by pressing Shift + Space

We are providing users with disambiguation of their search term in the first step. This is supported only when selecting the Freebase endpoint.

Figure 14: The user selects one of the disambiguations
Figure 15: The user selects one of the categories
Suggestions will be provided based on a selected category. These categories represent, in the case of Freebase, the commons section in the respective resource page.

Figure 16: Multiple suggestions can be chosen
Figure 17: The selected items are added as sub-nodes of the starting node, with links back to the resource page of the Linked Data provider

Figure 18: Other Linked Data providers can be chosen

We have added so far Freebase, Wikipedia, and MindMeister Public Maps as sources for Wunderkind suggestions. In the tests conducted to far, Freebase yields the best results, with Wikipedia coming close in quality.

3.2 Publishing Semantic Mind Maps in the LOD

Within the CODE project, it is highly desired to publish the extracted as well as user generated data in a Linked Open Data aware manner. Following this regard, an endpoint has been implemented for the
industrial partner MindMeister. To overcome the shortcoming of the previously mentioned maintenance issues, an automated service process has been implemented to trigger a periodic update of changed data.

In detail, the following information is stored in the according endpoint for MindMeister. As shown in other CODE deliverables, mind maps support the user to visually manage the extracted and generated data, e.g., produced by the CODE Query Wizard. Further, the encompassed information can be easily changed and shared with others. To follow the idea of Linked Open Data, the MindMeister API has been extended to dump each public mind map using RDF.

Exporting a mind map in RDF can be done in two ways. First, any public mind map can be accessed at:

http://www.mindmeister.com/[map_id].rdf

where [map_id] is the public ID of the specific mind map. Secondly, it can be downloaded as RDF via the MindMeister export dialog (see figure below).

![Figure 19: Export dialog in MindMeister with the option "Export as RDF file"](image)

The entire structure of the mind map is encoded in RDF and the given format is Notation 3 (N3) (Figure 20). In particular the following data is included:
Map:
- title
- description
- date of creation
- date of last update
- is shared

User (creator and collaborators):
- name
- user id
- url to channel page on MindMeister

Node in the map:
- parent node
- label
- creator
- date of creation
- date of last update
- note
- link
- images
- attachments
- connection to other nodes

Figure 20: Graphical view of a mind map (left) and corresponding RDF encoding (right)
Further information on the creation of the endpoint can be found in D3.1 and D3.3 accordingly. The endpoint can be accessed at:

http://zaire.dimis.fim.uni-passau.de:8080/bigdataMindmeister/

Currently, the input routines are up and running. After a short revision phase, both endpoints will be registered at the CKAN platform with their specific information.

3.3 From Mind Maps to Presentations

The capability of importing PDFs on MindMeister allows the automatic creation of mind maps from the table of content of an arbitrary scientific paper (Figure 21).

![Data flow of mind map creation from table of content of scientific papers](image)

Figure 21: Data flow of mind map creation from table of content of scientific papers

The PDF file is sent to MindMeister via an import dialog in the web interface. The next step includes the extraction of the table of content of the paper. This is done by the Enrichment Service. The input is the binary data of the PDF and the table of content is returned in JavaScript Object Notation (JSON) back to MindMeister.
The final step is made up of the creation of a mind map in the fashion that each header corresponds to a node in the mind map and its level represents the level of the node in the mind map.

Using MindMeister’s auto-create presentation function, the mind map can now be turned into an online slideshow at the click of a button (Figure 22). Researchers can use this functionality to quickly create an overview of a research paper and publish it online, in turn making it available as RDF linked data through the feature explained above.
4 42-data: A data centric question and answering portal

During the first year we analysed the potential of data marketplaces for research data. Our surveys in [Hollauf2013] showed that such a market place could be hardly driven by monetary incentives alone. Further, our analysis on existing market place (including those already shutdown) revealed that all successful data marketplaces place their data in a particular context. Through this most often very focused context, data gains value. For example, in 2010 factual stated that

“Factual is a platform where anyone can share and mash open data on any subject”\(^1\).

However, they narrowed down their scope in 2014, where factual is

“Factual is a location platform that enables personalized and contextually relevant mobile experiences by enriching mobile location signals with definitive global data.”\(^2\)

Based on this observation, the consortium developed the vision of a marketplace for empirical observations and statistical data in research-oriented processes. However, the challenge is still on how to engage users into such a marketplace, especially if a large fraction of users behaves out of an altruistic motivation.

When analysing today’s data portals we identified that they all have capabilities to manage and access open data sets, but not to consume and interact with data. They miss the opportunity in creating a community of interest around data and to engage into human interaction about interesting facts. These factors, however, have been the main success factors for some of the most successful websites today, like for example Facebook, Wikipedia, Stackexchange and Mendeley.

For example, the Digital Agenda EU Portal provides a large number of charts and datasets to analyse important indicators for Europe\(^3\). However, it does not allow people to share their findings on data with each other and discuss its interpretations. Hence, the context and meaning for the data is missing.

Also, finding interesting insights in data is a time consuming part. So not every user actually wants to engage in this finding process and needs the support from others to discover interesting patterns.

These observation lead to the development of a data-centric question and answer portal for empirical research data to allow social interactions to be build around Linked Open Data for research.

4.1 42-data: Data in context

42-data is a data-centric question and answering portal build around 4 main functions for its users, namely to

1. Answer research-related questions using (Linked) Open Data
2. Bookmark (Linked) Open Data found on the Web to your library
3. Discover interesting insights in (Linked) Open Data


\(^3\) [http://digital-agenda-data.eu/charts](http://digital-agenda-data.eu/charts)
4. Donate for impressing answers, questions, data as well as their provider

In this sense, 42-data forms a virtual meeting place for people interested in getting insights from open data sets with a focus on research-related data, and in sharing their insights with others. While it does not form a highly commercially-oriented market place, it allows acknowledging valuable insights through two main mechanisms: reputation and donations.

**Reputation** is a core driver on a number of major web platforms, like for example Stackexchange\(^4\), Wikipedia\(^5\) or Github\(^6\). The reputation does help users to improve their social and most often also their employment position. Hence it provides indirect monetary (and also non-monetary) value. With 42-data we provide a similar platform that Github provides for open source projects, but just for open data. Users can **gain reputation through analysing data sets and creating interesting insights.**

**Donations** form the basis for monetary incentives. Donations have been proven to successfully finance large web sites like Wikipedia, but are also applicable for small, one-person companies like for example small software-tool developers or nice-game developers. 42-data supports the creation of such monetary revenue chains by allowing people to donate for users, questions, answers or resources provided. Also, using donations for providing monetary incentives to 42-data users avoids cost-intensive processes like contracts, different payment forms etc. Hence, we can exploit the long-tail of micro-payments and evaluate the value contained in open data.

Besides market place-oriented aspects 42-data also offers services not yet available to the user. In particular:

- **Exploring (Linked) Open Data:** 42-data offers services for discovering relevant data sets in (Linked) Open Data. The balloon service [Bayerl2014] gathers necessary structural information of the Web of Data while the CODE Query Wizard [Hoefler2014] allows for easily searching the Linked Open Data cloud. Through exploring (Linked) Open Data our users should be empowered to aggregate and integrate interesting data. The results of both activities lead to a central storage of primary research-data in the form of data cubes maintained by the designated 42-data endpoint.

- **Data visualisation and analysis:** On the basis of the data cubes stored in the 42-data endpoint, the CODE Visualization Wizard [Hoefler2014] provides visual analysis functionality. This prototype offers interactive and context-dependent visualizations with a broad field of possible diagrams.

- **Data bookmarking:** Since the structure of 42-data is resource centric, a user is able to utilize the portal as bookmarking service for own data resources. Further, a data provider is in future able to register her/his own data endpoints.

- **Community-driven insight generation:** Through question asking and answering we hope to stipulate new discoveries and interesting findings based on open data. It is well known in academia that discoveries come out of intense communication processes. 42-data supports exactly those.

The given set of available services shows the applicability of the portal in the research domain. In future, it is envisioned to support research communities as well as scientific events (e.g., conferences or evaluation campaigns) by those workflows.

---

\(^4\) [http://www.stackexchange.com/](http://www.stackexchange.com/)


\(^6\) [https://github.com/](https://github.com/)
4.2 The donation and reputation concept

As already shown, question and answering portals have been highly successful in distilling the most valuable knowledge in particular domains. However, answers are based on opinions, rather than facts.

In today’s landscape of the Web, two different types of question and answering portals are observable: most of the domain independent question and answering portals suffer a slow uptake, such as the aforementioned Kasabi platform. In contrast to that, highly domain dependent and community-driven portals exist, such as StackOverflow7. Those portals exhibit a tremendous activity frequency between the participating users. Further, community portals are driven by sophisticated trust and reputation models to ensure high quality data. Those models vary from simple post processing of threads up to a complex rating model to infer questions and answers that expose high community impact. Besides already achieved benefits in such community portals, however they still lack in two major aspects:

- The first aspect is the integration of the data itself. Most portals simply link out to the given data silo inheriting the data portion solving the question. It is up to the user to get access to the data as well as to perform data integration to establish a basis for further analysis.
- The second aspect tackles the overall marketplace that drives the user’s motivation to interact with the portal by formulating a high impact question or answer or to curate an existing data.

42-data proposed by the CODE project addresses those issues of current data marketplaces. It integrates the CODE research prototypes and is therefore capable to offer a sophisticated end-to-end data integration and analysis workflow as shown above. Within this workflow already available trust and reputation models in combination with the proposed provenance chains enable a complete revenue value chain. Here, a donation-based revenue model enables a win-win situation for all participating peers. In this model, anybody who is interacting with the global question will be rewarded with a monetary benefit. A monetary incentive is present by donations. Those donations are realized by micro-payments processed by Paypal. In this regard every peer is able to donate for a specific donation sink in the portal. A donation sink can be a complete question, a specific answer, a person, a data set or a data provider.

4.3 Available resources in 42-data

As already mentioned, the structure of the portal is resource centric. Therefore a user has to be guided in the workflows to enable the integration of straightforward resources as well as resources with a high complexity. Within the portal, we distinguish the following resources:

Free-text resource is the most general resource in the portal. It exhibits an optional title and a description field. Here, a description can be formatted with a Markdown Editor to enable WYSIWYG formatting.

All further resources extend the free-text resource and store title and description accordingly.

Web resource can be any URL, which should be dereference-able. Figure 23 shows the integration of a Slideshare presentation inside an answer. For several scientific platforms, the discussed resources can be directly embedded. Those are dipity, mindmeister, Sketchfab, Slideshare, Speaker Deck, TED, and YouTube.

7 http://stackoverflow.com/
**Discussion Resource: Slides**

**Slides**

This slides at SlideShare seem to be incomplete. A complete version of my slides can be found [here](link).

---

**Figure 23: Integration of web resources in 42-data**

*Scientific paper* serves as perfect input for 42-data in two major ways: On the one hand, a paper is the direct outcome of one’s research work and therefore worthy to discuss. On the other hand, the services produced by WP2 [Kern2013] enable us to reverse engineer the paper to extract the primary research data available in tables. The outcome of this extraction is again a data cube. The exact identification of a paper is done on the basis of its Digital Object Identifier (DOI). Figure 24 illustrates the integration of scientific papers on the basis of Mendeley in the profile page of a user.

---

**Figure 24: Integration of Mendeley resources in the profile page of a 42-data user**
**SPARQL query** exposes very complex semantics and structure. The main advantage of bookmarking a data portion via its descriptive query is the up-to-date character of the data. Each time the surrounding website is rendered the query will be evaluated on the fly. Of course, expert users can only formulate SPARQL queries. This drawback is softened by the CODE Query Wizard that is capable of formulating queries in a Google-like behaviour and features a SPARQL export of the final query. Amongst different options, a user can include the SPARQL query to be shown in the answer and several kinds of visualisations, like for example a geospatial visualisation.

SPARQL resources aim at technically skilled users to lift semantic web resources to non-technical users and to reveal the answers hidden in Linked Open Data repositories. Figure 25 shows an answer including the countries with the highest percentage of Gross Domestic Product coming from agriculture. In this example, the following information is displayed in the answer: the endpoint hosting the data, the actual query and a visualization, here a Geo-based diagram.

![SPARQL query example](image)

**Figure 25: A SPARQL query resource in an answer**

**Data cube** is beside a SPARQL query the most complex resource in 42-data. On its basis, a data cube is a multidimensional data structure formulated in the RDF Data Cube Vocabulary.\(^8\) The portal further offers export interfaces, such as ARFF, to inject the data portions in data mining applications for further processing. As already mentioned, most workflows inside 42-data aim towards the creation of data cubes. The data cube resource itself can be embedded as (downloadable) raw data in the portal with a table-based visualization. An example integration of a data cube inside an answer is shown in Figure 26. Here, the data cube shows the funding of the University of Passau in completed FP7 projects.

\(^8\) [http://www.w3.org/TR/vocab-data-cube/](http://www.w3.org/TR/vocab-data-cube/)
Visualization enables visual analysis of data cubes, and supports the user by automating the visualization process. This means that the CODE Vis Wizard [Hoefler2014] automatically suggests any of the 10 currently available visualizations that are suitable for the provided data cube. Furthermore, the Vis Wizard (see Figure 27) automatically maps the data on the available visual channels of the chosen visualization. If the user wishes to adjust the mapping, this can be achieved with a few simple clicks.

Figure 27: The CODE Vis Wizard displays an interactive visual representation of the percentage of public services available online. Austria is selected in the left chart by the user and automatically highlighted in the right chart by the system.
Mind map is a powerful structure for summarization of information (cp., semantic mind maps). It is embedded in 42-data accordingly to be used in answers as well as questions. This integration is currently an on-going task.

The current set of resources is not closed. If new usage domains arise that require further resources, they can be added accordingly.

4.4 Integrated workflows in 42-data

The main usage aspects of the 42-data portal are centred on resources. By this design consideration, it was possible to bring together all available prototypes within 42-data. In the following, the basic workflows are introduced.

4.4.1 Creation of data cubes on the basis of (Linked) Open Data

The creation of data cubes is the foundation for sophisticated data management and analysis techniques. In the discover page (see Figure 28) of the portal, an overview of the Balloon\textsuperscript{9} \cite{Bayerl2014} index is presented. This index shows the type of the reachable endpoints, the link to it and the associated amount of data cubes. Via this page, the two different ways for data cube creation can be triggered.

![Discover Data](Image)

**Figure 28: Discovery page of 42-data**

**Exploring (Linked) Open Data.** Although the concept of Linked Data has been increasing in popularity, easy-to-use interfaces to access and make sense of the actual data are still few and far between.

When it comes to working with data, many people know how to use spreadsheet applications, such as Microsoft Excel. In comparison, very few people know SPARQL, the W3C standard language to query Linked Data. The CODE Query Wizard\textsuperscript{10} provides a web-based interface that dramatically simplifies the process of displaying, accessing, filtering, exploring, and navigating the Linked Data that’s available through a SPARQL endpoint. The main innovation of the interface is that it turns the graph structure of Linked Data into tabular form and provides easy-to-use interaction possibilities by using metaphors and techniques that the end user is already familiar with.

\textsuperscript{9} \url{http://schlegel.github.io/balloon/}
\textsuperscript{10} CODE Query Wizard: \url{http://code.know-center.tugraz.at/search}
The CODE Query Wizard offers two entry points: A user can either initiate a keyword search over a Linked Data repository, or select any of the already available datasets, represented as RDF Data Cubes. In both cases, the CODE Query Wizard presents a table containing the results. The user can then select columns of interest and set filters to narrow down the displayed data. Additionally, the user can explore the data by “focusing” on an entity, or can aggregate a dataset to obtain a summary of the data. The result will be stored as a data cube in the 42-data endpoint.

**Lifting offline data.** The Open Data movement is still an on-going activity and there exist lot of legacy documents, which are not available in a Web-enabled representation or integrated in retrieval engines.

To overcome this task, the Data Extractor has been generated to lift data into a Linked Data compliant data cube representation. Within the generated process, several input formats are considered as shown in [Bayerl2014]. In summery, the two major usage scenarios are as follows: i) transformation of documents characterized by low schema complexity but large volumes (e.g., CSV) and ii) aggregated facts like tables with low volume, but high schema complexity leading to a huge integration effort. In the 42-data platform, the Data Extractor has been integrated to manage the transformation process for large CSV or Excel Files as well as tables extracted by WP2 from research papers. Figure 30 shows the integration of the data Extractor in 42-data.
4.4.2 Creating questions and answers – a resource centric approach

The creation of discussions inside the 42-data portal is the most central workflow. In this sense, questions and answers can be generated in an easy fashioned way. Figure 31 introduces the screenshot of the question creation page. Here, the title and a description of a question are defined. Below the markdown preview window, the discussed resource can be added. It is possible to select one of the aforementioned resource types in a question. Tags, the amount of a guaranteed donation for question answering and the possibility to add a discussion feature are added as meta-information.

Figure 30: Integration of Data Extractor in 42-data

Figure 31: Question creation in 42-data
It is clear, that a question can be arbitrary complex in means of the attached resource. Along with this, an answer can be formulated on the basis of a sequence of resources. This is shown in Figure 32. In this example, the answer consists of a plain-text resource and an online resource. Further resources may be added accordingly.

Figure 32: Formulating an answer with several resources
4.4.3 Bookmarking resources for creating your own data library

The 42-data platform offers the possibility to be called by external services via an API. Following this, it is possible to inject resources directly into the system by external components. Currently, the following applications use this functionality:

- **CODE Query Wizard** exhibits three different ways of creating a resource inside the data portal: The first functionality is to store the selected data cube directly in the 42-data endpoint. Besides storing the concrete data portion, it is possible to store the SPARQL query to enable the dynamic evaluation inside the corresponding entity.
- **CODE Browser Extension** offers the possibility to crawl tables embedded in HTML tables and to forward it to the data portal. A syntactical merging step as pre-processing can be applied as well.

![Figure 33: Injection of external resources into 42-data](image-url)
## 5 Success Factors

In this section we report on the success factors achieved in Phase 2 along with the above prototypes. For each success factor we give a quantitative estimate and a qualitative description how it has been calculated.

<table>
<thead>
<tr>
<th>Success Factor</th>
<th>Goal Phase 2</th>
<th>Status</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithmic Annotations and Integrations per Day</td>
<td>400,000</td>
<td>700,000</td>
<td>Annotations include tables, figures, table of content and entities. The numbers are monitored by the enrichment service.</td>
</tr>
<tr>
<td>Accessible LOD Repositories</td>
<td>10</td>
<td>229</td>
<td>The number of repositories includes also repositories of bad service quality. Most of the 229 services are hardly usable due to slow response time, unreliable uptimes and non-fuzzy search capabilities. Currently 46 endpoints are offered via CODE Services, which provide more reliable performance. The numbers are monitored by the Balloon Service.</td>
</tr>
</tbody>
</table>
| Validated Annotations                                | 10,000       |        | The number of validated annotation is composed as follows:  
- Manual PDF annotations have been created in Phase II as test data for computer science extractors  
- We manually controlled 1000 PDFs and rejected wrong annotations of tables, figures and table of content.  
- Data Cubes consist of annotations and integrations to the LOD cloud. We validated 10 data cubes with approximately 100 annotations per cube as test data  
- We converted public mind-maps to RDF using the SKOS Format. Those mind-maps are available as SPARQL endpoints and since they are manually created they also form validated annotations. |
| Aggregated Data Sets                                 | 100          |        | The number of aggregated data sets is constituted as follows  
- The 400 data cubes have been created as part of the development process of the 42-data portal. The data cubes are very noisy and subject to data cleaning  
- Given the current enrichment and disambiguation service, we enriched all open-access publications available in Mendeley. The resulting 750k research papers have been annotated with tables, figures and table of content information.  
- We converted public MindMeister mind-maps to RDF using SKOS. Those mind-maps are available as SPARQL endpoints and since they are manually created they also form validated annotations. |
| Query API Calls from external users                  | 10,000       | None   | Our services have not been exposed to external users yet in a significant manner. Hence, we did not count the number of API calls as of now. For Phase |
### III we count the following numbers for the API Calls:
1. SPARQL calls to the 42-data endpoints (including Data Cubes and MindMeister Thesaurus)
2. Number of requested RDF Maps at MindMeister
3. Number of API calls accessing PDF Annotations at Mendeley

<table>
<thead>
<tr>
<th>External User/Domain Experts</th>
<th>200</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>We did not expose our services to public users. The external users counted have been external users like Mendeley advisors or students involved in testing our services. In Phase III we will expose all services to a large community. The following user counts will be measured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Number of Mendeley Desktop users using enriched PDFs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Number of MindMeister users using enriched mind-maps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Number of 42-data users</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Recommendations Accepted</th>
<th>10,000</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number has not been tracked yet due to insignificant exposure to external users. The number will be measured as</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Number of mind-map extensions through structured data (Wunderkind)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Number of disambiguation requests for table disambiguation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Number of requests to the Balloon service</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income to cost ratio</th>
<th>Negative</th>
<th>Negative</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Data Quality</th>
<th>Good</th>
<th>Good/Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>We conducted the following data quality estimates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- We manually evaluated the annotations of 1000 tables, figures, table of contents and entities through the Mendeley QA team. The result show, that tables, figures and table of contents are of high enough quality to be exposed to the public Mendeley users. However, despite acceptable quantitative evaluations, entities lack the necessary accuracy and completeness to be exposed to the average user</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- We conducted quantitative evaluations of enrichment and integration algorithms and achieved state-of-the-art precision and recall values (see D 2.2).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- We conducted standardization tests with our data cubes and evaluated their structural coherence. The cubes are structural coherence, but the semantic mapping has still to be improved</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6 References


