D3.2 - Provenance Schema

Summary: This deliverable highlights the CODE approach to model provenance information. Provenance information is omnipresent in all workflows between the available project prototypes. The main objective is to establish a full revision chain of data, which involves statements about the complete history of ownership. This information will be the foundation for further processes, such as marketplace concepts as well as query optimization.

<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>2013-04-30</td>
</tr>
<tr>
<td>Nature</td>
<td>O (Other)</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU (Public)</td>
</tr>
<tr>
<td>WP Lead Partner</td>
<td>University of Passau</td>
</tr>
<tr>
<td>Revision</td>
<td>final</td>
</tr>
<tr>
<td>Authors</td>
<td>Sebastian Bayer, Michael Granitzer, Manuel Grabowski, Kai Schlegel, Florian Stegmaier</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Statement of originality: This document contains original unpublished work except where clearly indicated otherwise.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
Commercially Empowered
Linked Open Data Ecosystems in Research

Project Officer & Project Coordinators

<table>
<thead>
<tr>
<th>Project Officer</th>
<th>Stefano Bertolo</th>
<th>European Commission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Directorate-General Communications Networks, Content and Technology (DG CONNECT)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unit Data Value Chain (G3) 10, Rue Robert Stümper</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2920 Luxembourg</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="mailto:stefano.bertolo@ec.europa.eu">stefano.bertolo@ec.europa.eu</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Project Coordinator</th>
<th>Stefanie Lindstaedt</th>
<th>Know-Center GmbH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inffeldgasse 13, 8010 Graz, Austria</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+43 316 873-30800 (phone)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+43 316 873-1030800 (fax)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="mailto:slind@know-center.at">slind@know-center.at</a></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scientific Coordinator</th>
<th>Michael Granitzer</th>
<th>University of Passau</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Innstrasse 33a, D-94032 Passau</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+49(0)851-509-3305</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="mailto:michael.granitzer@uni-passau.de">michael.granitzer@uni-passau.de</a></td>
<td></td>
</tr>
</tbody>
</table>

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</td>
<td>2013-03-07</td>
<td>Stegmaier</td>
<td>UNI PASSAU</td>
<td>Creation of Table of Contents</td>
</tr>
<tr>
<td>2</td>
<td>2013-03-25</td>
<td>Stegmaier</td>
<td>UNI PASSAU</td>
<td>Added introduction, workflows analysis, draft, defined data-centric provenance chains</td>
</tr>
<tr>
<td>3</td>
<td>2013-03-25</td>
<td>Granitzer</td>
<td>UNI PASSAU</td>
<td>Review of Workflows</td>
</tr>
<tr>
<td>4</td>
<td>2013-04-04</td>
<td>Stegmaier</td>
<td>UNI PASSAU</td>
<td>Revised section 2, Inserted Provenance Examples, added related work, description of ontology and conclusion</td>
</tr>
<tr>
<td>5</td>
<td>2013-04-07</td>
<td>Granitzer</td>
<td>UNI PASSAU</td>
<td>Review and Revision</td>
</tr>
<tr>
<td>6</td>
<td>2013-04-09</td>
<td>Stegmaier</td>
<td>UNI PASSAU</td>
<td>Integrated Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-------</td>
<td>----------</td>
<td>------------</td>
<td>--------------</td>
</tr>
<tr>
<td>7</td>
<td>2013-04-18</td>
<td>Maurer</td>
<td>Know-Center</td>
<td>Proof reading</td>
</tr>
<tr>
<td>Final</td>
<td>2013-04-30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table of Contents

1 Introduction ...........................................3

2 Analysis of Data Interaction Workflows .................4
   2.1 DATA CREATION ..................................6
   2.2 DATA TRANSFORMATION .........................8
   2.3 DATA CONSUMPTION ..............................8

3 Modelling Provenance Information in the CODE project 10
   3.1 RELATED WORK .................................10
   3.2 PROVENANCE ONTOLOGY .......................10
   3.3 SCENARIOS FOR EXAMPLE PROVENANCE CHAINS 14

4 Conclusion & Outlook ................................15

5 References ...........................................16

Appendix A: Provenance Data for Scenario 1 ............18
Appendix B: Provenance Data for Scenario 2 ............21
Appendix C: Provenance Data for Scenario 3 ............24
Appendix D: Provenance Data for Scenario 4 ............27
Appendix E: Provenance Data for Scenario 5 ............30
1 Introduction

Within the last ten years, the Web reinvented itself over and over again, which led from a more or less static and silo-based Web to an open Web of data, the so called Semantic Web\(^1\). Its vision is to provide an open-access, machine-readable and semantic description of content, mediated by ontologies. Following this, Linked Data [Bizer, 2009] is the de-facto standard to publish and interlink distributed data sets in the Web. At its core, Linked Data defines a set of rules on how to expose data and leverages the combination of Semantic Web best practices, e.g., RDF\(^2\).

However, the Linked Data cloud is mostly restricted to academic purposes due to unreliability of services and a lack of quality estimations of the accessible data. The vision of the CODE project\(^3\) is to improve this situation by creating a web-based, commercially oriented ecosystem for the Linked Science cloud, which is the part of the Linked Data cloud focusing on research data. This ecosystem offers a value-creation chain to increase the interaction between all peers, e.g., data vendors or analysts. In this sense, the integration of a marketplace leads on the one hand to crowd sourced data processing and on the other hand to sustainability. The original motivation behind the CODE project originated from obstacles of daily research work. When working on a specific research topic, the related work analysis is a crucial step. Unfortunately, this has to be done manually and usually is time consuming due to the following facts: First, experimental results and observations are locked in PDF documents, which are out of the box unstructured and not efficiently searchable. Second, there exist a large amount of conferences, workshops, etc. leading to a tremendous amount of published research data. Without doubt, the creation of a comprehensive overview over ongoing research activities is a cumbersome task. Moreover, these issues can lead to a completely wrong interpretation of the evolution of a research topic. Specifically for research on ad-hoc information retrieval, Armstrong et al. [Armstrong, 2009] discovered in an analysis of research papers issued within a decade that no significant progress has been achieved. These issues could be improved by the usage of services established by the CODE project.

Within the tackled use case, a central aim is to utilize data provenance along the central steps in the data lifecycle, e.g., creation, consumption and processing, along corresponding peers/users in order to enable monitoring and data quality estimations. As implied, data provenance is the foundation towards mature data processing and analysis, but has to be injected in various workflows to utilize its full potential. Therefore this deliverable focuses on:

- Analysis and specification of data interaction workflows inside the CODE project
- Definition of a Semantic Web aware provenance data model to annotate primary research data

The structure of this report is as follows: Section 2 analysis CODE workflows with a focus on gatherable provenance information. Section 3 gives insights into current activities in the domain of provenance ontologies and discusses the decision to use the PROV ontology in the CODE project. Section 4 concludes the deliverable.

Background knowledge in terms of data cubes as well as triplification of unstructured data as specified in [Bayerl, 2012] is postulated.

\(^1\) http://www.w3.org/2001/sw/
\(^2\) http://www.w3.org/RDF/
\(^3\) http://www.code-research.eu/
2 Analysis of Data Interaction Workflows

The different data interaction workflows in the CODE project depend on the utilized data sources. Research data is made available to the research society in various ways, e.g., through digital libraries or specific websites. Table 1 summarizes four data sources that are taken into account in the aforementioned usage scenario.

<table>
<thead>
<tr>
<th>Type</th>
<th>Data Set Description</th>
<th>Data Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research paper</td>
<td>PDF documents</td>
<td>Aggregated facts like tables, figures or textual patterns. Low volume, but high integration effort.</td>
</tr>
<tr>
<td>Primary research data</td>
<td>Evaluation data of research campaigns available in a spreadsheet like format (1. normal form) or via Web-APIs</td>
<td>Data generated by mostly automated means. Large volumes, low schema complexity.</td>
</tr>
<tr>
<td>Retrievable data</td>
<td>Linked Open Data endpoints</td>
<td>Semantically rich, interconnected data sets. Large volumes, hard to query (technically and from a usability point of view). Mostly background knowledge.</td>
</tr>
<tr>
<td>Embedded data</td>
<td>Microdata(^4), microformat(^5) and RDFa(^6)</td>
<td>Semantically rich, but distributed data. Less of interest.</td>
</tr>
</tbody>
</table>

Table 1: Data sources for primary research data in scope

Research papers are a valuable source of state-of-the-art knowledge which are mostly stored in digital libraries reaching an amount of several Terabytes. Apart from the overall storage, the actual size of a single PDF document does not exceed a few Megabytes. The main task is to extract meaningful, encapsulated information such as facts, table of contents, figures and - most important - tables carrying the actual evaluation results. The present diversity of extracted data leads to a high integration effort for a unified storage. In contrast to that, primary research data is released in a more data centric form, such as table-based data. This kind of data is mostly issued by (periodic) evaluation campaigns or computing challenges. Famous examples are the CLEF initiative\(^7\) focusing on the promotion of research, innovation, and development of information retrieval systems. The outcome of such activities is thousands of raw data points stored in Excel sheets. Here, the volume of the data is most likely very large but defined by a specific schema with less complexity than PDF documents. Both data sources share an unstructured nature due to missing semantics on the schema and data level. To overcome this issue, the two remaining data sources of Table 1 are utilized. In this light, Linked Open Data endpoints serve as source for retrievable data, such as DBPedia\(^8\) or PubMed\(^9\). On the one hand, these endpoints expose their data following the 5 star open data rule\(^10\) meaning

\(^4\) http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html
\(^5\) http://www.microformats.org/
\(^6\) http://www.rdfa.info/
\(^7\) http://www.clef-initiative.eu/
\(^8\) http://www.dbpedia.org/
\(^9\) http://pubmed.bio2rdf.org/
\(^10\) http://www.5stardata.info/
the data is openly available, annotated with clear semantics and interconnected in the distributed Linked Open Data cloud. On the other hand, due to its distributed nature, efficient federated retrieval is a hard task. The last data source mentioned is embedded data, meaning content of websites semantically annotated with microdata, microformat or RDFa. This information can be embedded table-based primary research data or auxiliary information, such as biographic data of a person.

In order to process the set of different data sources and to enable user interaction, five specific user interface prototypes have been developed:

- **Data Extractor** [Bayerl, 2012] is the central component for data cube creation and triplification of unstructured data. The data extractor converts arbitrarily formatted tabular data into RDF data cubes through semi-automatic means. It processes PDF documents using WP2 services, table-based data (SPARQL Query results and primary research data in the form of spreadsheets) as well as embedded data.

- **Mendeley Desktop**\(^{11}\) is able to process PDF documents stored in the Mendeley library in terms of fact and table extraction, table of content analysis and disambiguation of concepts by using WP2 services. It serves as integration point for several data sources and as link to Mendeley’s large user base.

- **Query Wizard** [Mutlu, 2012] enables a Google-like discovery of linked data endpoints. During discovery, the data cubes are being created or already existent cubes are enlarged/enriched. The query Wizard uses federated querying techniques in order to access a large body of the Linked Open Data cloud in an easy usable way.

- **Visual Analytics** [Mutlu, 2012] serves as the data consumer and offers possibilities for visual analytics over data cubes. It integrates different kinds of visualisations, which can be organized in a hierarchical manner through MeisterLabs MindMapping\(^{12}\) service.

---


\(^{12}\) [http://www.mindmeister.com/](http://www.mindmeister.com/)
All of those prototypes are using central CODE services, such as the data cube storage or the query federator. Figure 1 distinguishes the following three fundamental workflows and corresponding CODE services from an user perspective: *data creation*, *data transformation* and *data consumption*. Those workflows strongly differ in the data interaction level leading to specific provenance chains. Every provenance chain can be categorised according to the functional element as depicted in Figure 1 and the data source, as depicted in Table 1.

### 2.1 Data Creation

The most essential workflow is data creation to establish a Semantic Web aware storage layer with data centric retrieval functionalities comparable to data warehouses. Data creation happens either from

(i) Research papers like PDF documents  
(ii) Primary research data sources like XLS/CSV/HTML spreadsheets  
(iii) Retrievable data like Linked Open Data endpoints (RDF)  

Figure 2 shows the abstract provenance chain for data creation for all data sources in scope. Irrespective of the underlying prototypes, data creation from (i) and (ii) are covered through the same chain, while (iii) follows a slightly different one. The three variants will be described below from a data centric point of view with respect to the distinguished chains.

**Figure 2: Provenance chain for data creation**

**(i) & (ii) Data creation from research papers and primary research data**  
As indicated in Figure 1 the Mendeley Desktop and the Data Extractor are entry points for data creation in terms of unstructured data sources. Inside both prototypes, a user authentication is present enabling the collection of user information, which is interacting with the data. Here, a difference has to be made between the *producers* of data, e.g., the author of a processed paper, or the *integrator*, who is in charge of the actual data lifting process.

The first step towards data lifting is a source specific pre-processing:

- Due to the possible large volume of XLS/CSV data and the rather low schema complexity, the data is loaded by specific API calls into the Data Extractor prototype.  
- PDF documents will be analysed by WP2 services to extract facts and tables. The entry point for this operation can be either the Mendeley desktop or the Data Extractor.  
- A browser widget is utilized to unhinge tables embedded in webpages.
The results of these pre-processing steps are being transformed into a unified pivot table-abstraction format. The annotated tables are taken over by the Data Extractor prototype where the following steps happen. After preparation, a semi-automatic data refinement and enrichment on the annotated tables takes place. For data cleansing and patching transformation mistakes the user is able to make use of standard table manipulation functions, like edit cells or exclude columns/rows. To allow a high-level analytical retrieval over the exposed data the semantic structure of the table data has to be defined. In general, this structure is characterized by a set of dimensions and the observation values. To enable the enriched data for statistical analysis, the dimensions have to be classified by a measure type, e.g., nominal or ordinal. In addition, the semantic concept disambiguation service of WP2 is triggered for each table cell in order to interlink the primary research data to endpoints of the linked data cloud. The last step of the data lifting is the conversion of the semantically enriched tables to data cubes following the W3C RDF Data Cube Vocabulary specification. Finally, the actual storage of the data cube ends the provenance chain.

(iii) Data creation from Linked Open Data endpoints

In order to create data cubes from Linked Open Data endpoints the Query Wizard must be considered as entry point. An analyst or an integrator is able to retrieve portions of RDF data in a federated way from Linked Open Data endpoints. In this sense, the Query Wizard uses WP3s query federator, as illustrated in Figure 1. The RDF data portion can be manipulated and enriched by the user and in the end transferred to the Data Extractor. The annotated table is already semantically enriched with semantic concepts and therefore the disambiguation step is obsolete. In addition to the annotated table which is already constructed while browsing in the Query Wizard, user information as well as the actual SPARQL query is shipped to the Data Extractor. The remaining steps correlate to the description of (i).

Both variants of this provenance chain expose specific activities. Throughout all of these activities, crucial information always depicts the data source and the service or the peer enforced it. This information is summarized in Table 2.

<table>
<thead>
<tr>
<th>Activity concerning data</th>
<th>Participating Peers</th>
<th>User information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraction</td>
<td>Analyst</td>
<td>Mendeley or Mindmeister profiles</td>
</tr>
<tr>
<td>Selection</td>
<td>Integrator</td>
<td>External information embedded in HTML</td>
</tr>
<tr>
<td>Manipulation</td>
<td>Producer</td>
<td></td>
</tr>
<tr>
<td>Semantic enrichment</td>
<td>Software Agent (e.g., the prototype)</td>
<td></td>
</tr>
<tr>
<td>(E.g., disambiguation)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Available provenance information during data creation

---

13 See [Bayerl, 2013] for further details on the internal table format.

14 http://www.w3.org/TR/vocab-data-cube/

15 The profile information is gathered utilizing OAuth secure authentication.

16 External information is gathered using tools developed by getschema.org.
2.2 Data Transformation

In contrast to data creation, data transformation refines already existent data cubes to discover new correlations or to clean existing cubes. This is achieved by extending a data cube with (semantically linked) additional information available in the Linked Open Data cloud.

Figure 3 defines the associated provenance chain with the Query Wizard as entry point. As first step, the Query Wizard loads a data cube\(^\text{17}\) from the storage layer and displays it to an analyst. The user is now able to manipulate the available data cube and to enrich it with further RDF data collected from the Linked Open Data cloud. A user-driven Linked Open Data discovery is the foundation for this step, enabled by the CODE query federation services. While data browsing, a new data cube is generated and can be stored following data creation step (iii).

![Figure 3: Data centric provenance chain for data transformation](image)

Since the data transformation completely embeds the data creation step (iii), the set available provenance information defined in Table 2 is valid for data transformation as well. Furthermore, the data transformation provenance chain does not expend this set.

2.3 Data Consumption

The last provenance chain deals with the usage of data cubes. Data cubes can be loaded either from the Linked Open Data cloud or from the CUBE data cube storage layer. In the Visual Analytics prototype, a consumer analyses the data and creates visualizations on the basis of broad scope of charting tools.

![Figure 4: Data centric provenance chain for data consumption](image)

The data provenance chain is a read-only workflow. The information that is interacting with the data is needed as a foundation for billing mechanisms of the marketplace. It is envisioned to equip the Visual Analytics prototype with user information on the basis of Mendeley and Mindmeister profiles.

\(^{17}\) Note: The loaded data cube has already provenance information linked to it from the data creation step.
<table>
<thead>
<tr>
<th>Activity concerning data</th>
<th>Participating Peers</th>
<th>User information</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Consumption</td>
<td>• Consumer</td>
<td>• Mendeley or Mindmeister profiles</td>
</tr>
<tr>
<td></td>
<td>• Software Agent (e.g., the prototype)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Available provenance information during data consumption
3 Modelling Provenance Information in the CODE project

As shown in the previous chapter, provenance information is omnipresent in the CODE workflows. This section will introduce the PROV-O ontology, which is used in the CODE project to model provenance information in a Semantic Web aware manner.

3.1 Related Work

Data Provenance refers to the documented history and origin of data objects and has already gained wider attention before the rise of the Semantic Web. Provenance management and data lifecycle information are issues in the overall IT-landscape [Moreau, 2008a] [Freire, 2008] for example in the field of databases [Buneman, 2007] or process documentation in healthcare management systems [Kifor, 2006]. Beyond that, the characteristics of the Semantic Web, e.g., its open nature and many autonomous sources, make provenance information into a very important concept in the Web of Data in terms of quality assurance, trust and proof reasoning.

In particular, the Linked Data Cloud is accompanied by publishing data throughout several detached, autonomous but interlinked endpoints. This circumstance basically implies that published data could be incomplete, semantically inconsistent or contradicting. Also at the interaction level, one is not aware of a complete processing chain. This characteristic of openness and the thus resulting complexity prevents to take the entire knowledge of the Web of Data into account. However, if the real world is considered, one does not believe in all information and consequently disregards it. Our real-live trust is for example based on social aspects, empirical facts, recommendations or reputation. Transferring this to the linked data cloud means, that using trust and confidence in some portions of data could reduce the reasoning complexity and increase the result quality due to pruning of untrustworthy or even out-dated parts of the knowledge.

Following this, one way to integrate trust into the Web of Data is provenance. For simple metadata provenance information the Dublin Core vocabulary is sufficient. The Open Provenance Model OPM [Moreau, 2008b] defines a more sophisticated provenance model and is the result of several Provenance Challenge series, which was initiated in May 2006. Standardization bodies also addressed the need of the community for a unified provenance description schema. In this light the W3C Provenance WG recently issued the PROV ontology [Lebo, 2013]. Both models specify a pragmatic approach to provide a domain-independent interchange core model and basically distinguish between three provenance participants: artefacts/entity, processes/activity and agents.

To be more suitable for provenance on the Semantic Web, the Provenance Vocabulary was introduced as a domain specific extension of PROV-O. Besides pure provenance annotation, the querying of these information [Dividino, 2009] [Hartig, 2009] and consideration of provenance in the query execution (e.g. pruning and filtering of untrusted data or trust aware inference) are consequent topics [Ding, 2005] [Golbeck, 2008].

3.2 Provenance Ontology

The aforementioned discussion clearly states that provenance refers to information about the origin

---

18 http://dublincore.org/documents/dcmi-terms/
19 http://twiki.ipaw.info/bin/view/Challenge/
20 http://www.w3.org/2011/prov/
21 http://trdf.sourceforge.net/provenance/ns.html
or the source of data. This includes records about involved individuals, timestamps or process circumstances as already derived in Section 2 for each of the three CODE workflows. The main objective is to establish a full revision chain of data, which involves statements about the complete history of ownership. In order to fulfil this modelling task, the recently issued W3C PROV ontology is used. The usage of open standards is preferred to enable interoperability between different peers and to foster a wide adoption of the collected data.

The PROV ontology is of course generated for general needs and therefore offering a broad semantic expressiveness. A Semantic Web aware version of the ontology can be found online: http://www.w3.org/ns/prov-o. When looking at the needs of the CODE project, only a subset of the overall ontology is needed. By using only a subset without applied changes of the ontology, the provenance information is interoperable and still compliant to the standard. In the following, only this subset will be specified and the remaining parts of the official ontology will be ignored.

At its core, the ontology differentiates three main classes in a provenance chain:

- An entity describes the objects whose provenance should be specified or is source for another entity.
- How the entity was created, changed or extracted is described by an action.
- Finally, the agent point out who is responsible for the action.

The correlation of those main objects is visualized in Figure 5 and presents the utilized subset of the PROV-O data model (without namespace declaration due to space constraints). In fact, the ontology overlaps with Dublin Core in some parts (e.g., direct equivalents), but offer great extensions for more complex provenance information and provenance chaining.

![Figure 5: Simplified W3C PROV-O data model](image)

The properties interconnecting the classes carry the following semantics:

- wasDerivedFrom specifies a data transformation from a representation of an entity to another one.
- wasGeneratedBy identifies the actual creation of a new entity by a specific activity.
- used is the property opposite to wasGeneratedBy.
- wasAttributedTo is the link form an activity to an agent.
- wasAssociatedWith indicates that an agent had a specific role in an activity. Therefore it carries more semantic information as wasAttributedTo.
- wasInfluencedBy describes the influence of an agent on an activity.

In addition to the given classes and properties, specific controlled vocabularies / schemas or a more fine-grained differentiation of a class can be used for extension. For the first time, Semantic Web aware ontologies, such as FOAF for the definition of persons interacting with the data or Dublin Core, are suited to more general description tasks. At the class level, Person and SoftwareAgent are used as subclasses of Agent to clearly differentiate between a real person and the prototype in use.
To describe the CODE provenance chains with clear semantics, the CODE Data Provenance Vocabulary has been created as recommended by best practise. This vocabulary is not closed, nor does it pretend to be exhaustive. An overview of the current vocabulary properties can be found in Table 4 and Table 5. The properties have been defined in the following namespace:

http://code-research.eu/ontology/code-prov-vocabulary

Table 4: CODE Data Provenance Vocabulary for Participating Peers

<table>
<thead>
<tr>
<th>Participating Peer</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyst</td>
<td>Analyst</td>
</tr>
<tr>
<td>Consumer</td>
<td>Consumer</td>
</tr>
<tr>
<td>Integrator</td>
<td>Integrator</td>
</tr>
<tr>
<td>Producer</td>
<td>Producer</td>
</tr>
<tr>
<td>Data Extractor</td>
<td>DataExtractor</td>
</tr>
<tr>
<td>Mendeley Desktop</td>
<td>MendeleyDesktop</td>
</tr>
<tr>
<td>Query Wizard</td>
<td>QueryWizard</td>
</tr>
<tr>
<td>Visual Analytics</td>
<td>Visualizer</td>
</tr>
</tbody>
</table>

Table 4: CODE Data Provenance Vocabulary for Participating Peers

<table>
<thead>
<tr>
<th>Property</th>
<th>Affected provenance Chain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>annotatedTable</td>
<td>Data Creation (iii), Transformation</td>
<td><strong>Entity.</strong> Specifies a semantically enriched table ready to be stored as data cube by the Data Extractor.</td>
</tr>
<tr>
<td>classifiedAndInterlinkedData</td>
<td>Data Creation (i, ii)</td>
<td><strong>Entity.</strong> Specifies the result of the disambiguation step.</td>
</tr>
<tr>
<td>conversionToDataCube</td>
<td>Data Creation (i, ii, iii), Data Transformation</td>
<td><strong>Activity.</strong> The process of transforming the annotated table into a data cube.</td>
</tr>
<tr>
<td>dataCubeToBeRefined</td>
<td>Data Transformation</td>
<td><strong>Entity.</strong> Specifies an initial data cube that will be enriched.</td>
</tr>
<tr>
<td>dataCubeToBeDisplayed</td>
<td>Data Consumption</td>
<td><strong>Entity.</strong> Specifies a data cube that should be displayed.</td>
</tr>
<tr>
<td>defineAndClassifyData</td>
<td>Data Creation (i, ii)</td>
<td><strong>Activity.</strong> The process of selecting portions of the extracted data.</td>
</tr>
<tr>
<td>finalDataCube</td>
<td>Data Creation (i, ii, iii), Data Transformation</td>
<td><strong>Entity.</strong> Specifies the created data cube and therefore the end of a provenance chain.</td>
</tr>
<tr>
<td>Concept</td>
<td>Category</td>
<td>Type</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td>interlinkDataToEndpoints</td>
<td>Data Creation (i, ii)</td>
<td>Activity</td>
</tr>
<tr>
<td>refineAndCleanData</td>
<td>Data Creation (i, ii)</td>
<td>Activity</td>
</tr>
<tr>
<td>refinedAndCleanedData</td>
<td>Data Creation (i, ii)</td>
<td>Entity</td>
</tr>
<tr>
<td>retrieveRDF</td>
<td>Data Creation (iii), Data Transformation</td>
<td>Activity</td>
</tr>
<tr>
<td>retrievedRDF</td>
<td>Data Creation (iii), Data Transformation</td>
<td>Entity</td>
</tr>
<tr>
<td>manipulateAndEnrichRDF</td>
<td>Data Creation (iii), Data Transformation</td>
<td>Activity</td>
</tr>
<tr>
<td>manipulatedAndEnrichedRDF</td>
<td>Data Creation (iii), Data Transformation</td>
<td>Entity</td>
</tr>
<tr>
<td>pivotTableAbstraction</td>
<td>Data Creation (i, ii, iii), Data Transformation</td>
<td>Entity</td>
</tr>
<tr>
<td>transformIntoPivotFormat</td>
<td>Data Creation (i, ii, iii), Data Transformation</td>
<td>Activity</td>
</tr>
<tr>
<td>transferRDFToDataExtractor</td>
<td>Data Creation (ii, iii), Data Transformation</td>
<td>Activity</td>
</tr>
<tr>
<td>transferEnrichedDataCubeToDataExtractor</td>
<td>Data transformation</td>
<td>Activity</td>
</tr>
<tr>
<td>uploadPaper</td>
<td>Data Creation (i, ii)</td>
<td>Activity</td>
</tr>
<tr>
<td>visualizationMethod</td>
<td>Data Consumption</td>
<td>Activity</td>
</tr>
<tr>
<td>visualizeCube</td>
<td>Data Consumption</td>
<td>Entity</td>
</tr>
</tbody>
</table>

Table 5: CODE Data Provenance Vocabulary for Data Interaction
3.3 Scenarios for Example Provenance Chains

After the specification of workflows and the selection of an appropriate vocabulary to model the present provenance information, this section shows its applicability by the following five scenarios:

**Scenario 1:** In the first scenario, the integrator Max Muster adds the paper “A fantastic algorithm” written by John Doe to his Mendeley desktop. The Mendeley desktop extracts table-based data and hands it over to the Data Extractor. Max Muster runs through the manual processing steps of the Data Extractor and finally saves the data cube.

**Scenario 2:** Within this scenario, the same procedures are applied as in scenario 1. The only difference is the entry point. Instead of the Mendeley desktop, Max Muster uploads the paper “A fantastic algorithm” written by John Doe as PDF document directly to the Data Extractor in this scenario.

**Scenario 3:** The integrator Max Muster is going to create a new data cube from interesting data portions openly available in the Linked Open Data cloud. The data portions are located in two Linked Open Data endpoints and will be retrieved with a single query by the Query Wizard. After manipulation, the desired data will be transmitted to the Data Extractor and further processed.

**Scenario 4:** The analyst Max Muster wants to refine an existent data cube. He therefore loads the data cube into the wizard and extends the data cube with information located in two external Linked Data repositories. After manipulation, the data cube is passed to the Data Extractor and saved.

**Scenario 5:** The consumer Max Muster is loading a data cube and produces a chart in the Visual Analytics prototype.

Obviously, these five showcases directly correlate with the identified provenance chains. During every processing step, provenance information is recorded and stored. The complete provenance description of each scenario can be found in the corresponding appendix noted in Notation3\(^2\) format.

\(^2\) [http://www.w3.org/TeamSubmission/n3/]
4 Conclusion & Outlook

A main aim of the CODE project is the establishment of a marketplace to bootstrap the data ecosystem. A foundation for a marketplace is trust and one way towards trust is provenance. This deliverable presented an analysis of user-centric workflows between all prototypes to gain deeper understanding of available information of users and processes in terms of data interaction. On the basis of this analysis data-driven provenance chains have been derived to establish a full revision chain of data. To enable a persistent storage of gathered provenance data the W3C PROV ontology has been selected for modelling tasks.

In the current stage of the project, provenance data is only gathered through the workflows and is finally stored. In terms of provenance, year two of CODE will focus on the injection of this information in further services and processes, such as marketplace concepts and query optimization.
5 References


Appendix A: Provenance Data for Scenario 1

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix code-prov: <http://code-research.eu/ontology/provenance/> .
@prefix : <http://example.com#> .

:maxMuster
  a prov:Person;
  a code-prov:Integrator;
  foaf:givenName "Max Muster"^^xsd:string;
  foaf:mbox <mailto:max@muster.org>;
  foaf:homePage <http://max.muster.org>;
.

:johnDoe
  a prov:Person;
  a code-prov:Producer;
  foaf:givenName "John Doe"^^xsd:string;
  foaf:mbox <mailto:john@doe.org>;
  foaf:homePage <http://john.doe.org>;
.

code-prov:MendeleyDesktop
  a prov:SoftwareAgent;
.

:paper
  a prov:Entity;
  prov:wasAttributedTo :johnDoe;
  dcterms:title "A fantastic algorithm"^^xsd:string;
.

code-prov:uploadPaper
  a prov:Activity;
  prov:used :paper;
  prov:wasAssociatedWith code-prov:MendeleyDesktop;
  prov:wasInfluencedBy :maxMuster;
.
code-prov:DataExtractor
   a prov:SoftwareAgent;
.

code-prov:transformIntoPivotFormat
   a prov:Activity;
   prov:used :paper;
   prov:wasAssociatedWith code-prov:DataExtractor;
   prov:wasInfluencedBy :maxMuster;
.

code-prov:pivotTableAbstraction
   a prov:Entity;
   prov:wasDerivedFrom :paper;
   prov:wasGeneratedBy code-prov:uploadPaper, code-prov:transformIntoPivotFormat;
.

code-prov:refineAndCleanData
   a prov:Activity;
   prov:used code-prov:pivotTableAbstraction;
   prov:wasAssociatedWith code-prov:DataExtractor;
   prov:wasInfluencedBy :maxMuster;
.

code-prov:refinedAndCleanedData
   a prov:Entity;
   prov:wasDerivedFrom code-prov:pivotTableAbstraction;
   prov:wasGeneratedBy code-prov:refineAndCleanData;
.

code-prov:defineAndClassifyData
   a prov:Activity;
   prov:used code-prov:refinedAndCleanedData;
   prov:wasAssociatedWith code-prov:DataExtractor;
   prov:wasInfluencedBy :maxMuster;
.

code-prov:interlinkDataToEndpoints
   a prov:Activity;
   prov:used code-prov:refinedAndCleanedData;
   prov:wasAssociatedWith code-prov:DataExtractor;
   prov:wasInfluencedBy :maxMuster;
.
```

code-prov:classifiedAndInterlinkedData
   a prov:Entity;
   prov:wasDerivedFrom code-prov:refinedAndCleanedData;
   prov:wasGeneratedBy code-prov:defineAndClassifyData, code-prov:interlinkDataToEndpoints;
   .

code-prov:conversionToDataCube
   a prov:Activity;
   prov:used code-prov:classifiedAndInterlinkedData;
   prov:wasAssociatedWith code-prov:DataExtractor;
   prov:wasInfluencedBy :maxMuster;
   .

code-prov:finalDataCube
   a prov:Entity;
   prov:wasDerivedFrom code-prov:classifiedAndInterlinkedData;
   prov:wasGeneratedBy code-prov:conversionToDataCube;
   .
```
Appendix B: Provenance Data for Scenario 2

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix code-prov: <http://code-research.eu/ontology/provenance/> .
@prefix : <http://example.com#> .

:maxMuster
  a prov:Person;
  a code-prov:Integrator;
  foaf:givenName "Max Muster"^^xsd:string;
  foaf:mbox <mailto:max@muster.org>;
  foaf:homepage <http://max.muster.org>;
.

:johnDoe
  a prov:Person;
  a code-prov:Producer
  foaf:givenName "John Doe"^^xsd:string;
  foaf:mbox <mailto:john@doe.org>;
  foaf:homepage <http://john.doe.org>;
.

:paper
  a prov:Entity;
  prov:wasAttributedTo :johnDoe;
  dcterms:title "A fantastic algorithm"^^xsd:string;
.

code-prov:uploadPaper
  a prov:Activity;
  prov:used :paper;
  prov:wasAssociatedWith :maxMuster;
.

code-prov:DataExtractor
  a prov:SoftwareAgent;
code-prov:transformIntoPivotFormat
  a prov:Activity;
  prov:used :paper;
  prov:wasAssociatedWith code-prov:DataExtractor;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:pivotTableAbstraction
  a prov:Entity;
  prov:wasDerivedFrom :paper;
  prov:wasGeneratedBy code-prov:uploadPaper, code-prov:transformIntoPivotFormat;
.

code-prov:refineAndCleanData
  a prov:Activity;
  prov:used code-prov:pivotTableAbstraction;
  prov:wasAssociatedWith code-prov:DataExtractor;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:refinedAndCleanedData
  a prov:Entity;
  prov:wasDerivedFrom code-prov:pivotTableAbstraction;
  prov:wasGeneratedBy code-prov:refineAndCleanData;
.

code-prov:defineAndClassifyData
  a prov:Activity;
  prov:used code-prov:refinedAndCleanedData;
  prov:wasAssociatedWith code-prov:DataExtractor;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:interlinkDataToEndpoints
  a prov:Activity;
  prov:used code-prov:refinedAndCleanedData;
  prov:wasAssociatedWith code-prov:DataExtractor;
  prov:wasInfluencedBy :maxMuster;
.
Commercially Empowered
Linked Open Data Ecosystems in Research

D3.2 – Provenance Schema
Date: 2013-04-29

```prolog
code-prov:classifiedAndInterlinkedData
  a prov:Entity;
  prov:wasDerivedFrom code-prov:refinedAndCleanedData;
  prov:wasGeneratedBy :code-prov:defineAndClassifyData, code-prov:interlinkDataToEndpoints;
.

code-prov:conversionToDataCube
  a prov:Activity;
  prov:used code-prov:classifiedAndInterlinkedData;
  prov:wasAssociatedWith code-prov:DataExtractor;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:finalDataCube
  a prov:Entity;
  prov:wasDerivedFrom code-prov:classifiedAndInterlinkedData;
  prov:wasGeneratedBy code-prov:conversionToDataCube;
.
```
Appendix C: Provenance Data for Scenario 3

```xml
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix code-prov: <http://code-research.eu/ontology/provenance/> .
@prefix : <http://example.com#> .

:maxMuster
  a prov:Person;
  a code-prov:Integrator;
  foaf:givenName "Max Muster"^^xsd:string;
  foaf:mbox <mailto:max@muster.org>;
  foaf:homepage <http://max.muster.org>;
.

:linkedOpenDataEndpoint1
  a prov:SoftwareAgent;
  a code-prov:Producer
dcterms:identifier "http://my/endpoint1/"^^xsd:string;
.

:linkedOpenDataEndpoint2
  a prov:SoftwareAgent;
  a code-prov:Producer
dcterms:identifier "http://my/endpoint2/"^^xsd:string;
.

:mySparqlQuery
  a prov:Entity;
dcterms:identifier "actual query string"^^xsd:string;
.

code-prov:QueryWizard
  a prov:SoftwareAgent;
```
code-prov:retrieveRDF
  a prov:Activity;
  prov:used :linkedOpenDataEndpoint1, :linkedOpenDataEndpoint2, :mySparqlQuery;
  prov:wasAssociatedWith code-prov:QueryWizard;
  prov:wasInfluencedBy :maxiMuster;
  .

code-prov:retrievedRDF
  a prov:Entity;
  prov:wasDerivedFrom :linkedOpenDataEndpoint1, :linkedOpenDataEndpoint2, :mySparqlQuery;
  prov:wasGeneratedBy code-prov:retrieveRDF;
  .

code-prov:manipulateAndEnrichRDF
  a prov:Activity;
  prov:used code-prov:retrievedRDF;
  prov:wasAssociatedWith code-prov:QueryWizard;
  prov:wasInfluencedBy :maxiMuster;
  .

code-prov:manipulatedAndEnrichedRDF
  a prov:Entity;
  prov:wasDerivedFrom code-prov:retrievedRDF;
  prov:wasGeneratedBy code-prov:manipulateAndEnrichRDF;
  .

code-prov:DataExtractor
  a prov:SoftwareAgent;
  .

code-prov:transferRDFToDataExtractor
  a prov:Activity;
  prov:used code-prov:againManipulatedAndEnrichedRDF;
  prov:wasAssociatedWith code-prov:QueryWizard, code-prov:DataExtractor;
  prov:wasInfluencedBy :maxiMuster;
  .

code-prov:annotatedTable
  a prov:Entity;
  prov:wasDerivedFrom code-prov:manipulatedAndEnrichedRDF;
  prov:wasGeneratedBy code-prov:transferRDFToDataExtractor;
  .
code-prov:conversionToDataCube
  a prov:Activity;
  prov:used code-prov:annotatedTable;
  prov:wasAssociatedWith code-prov:DataExtractor;
  prov:wasInfluencedBy :maxiMuster;
.

:finalDataCube
  a prov:Entity;
  prov:wasDerivedFrom code-prov:annotatedTable;
  prov:wasGeneratedBy code-prov:conversionToDataCube;
.
Appendix D: Provenance Data for Scenario 4

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix code-prov: <http://code-research.eu/ontology/provenance/> .
@prefix : <http://example.com#> .

:maxMuster
  a prov:Person;
  a code-prov:Integrator
  foaf:givenName "Max Muster"^^xsd:string;
  foaf:mbox <mailto:max@muster.org>;
  foaf:homepage <http://max.muster.org>;
.

:linkedOpenDataEndpoint1
  a prov:SoftwareAgent;
  a code-prov:Producer;
  dcterms:identifier "http://my/endpoint1/"^^xsd:string;
.
:linkedOpenDataEndpoint2
  a prov:SoftwareAgent;
  a code-prov:Producer;
  dcterms:identifier "http://my/endpoint2/"^^xsd:string;
.
:mySparqlQuery
  a prov:Entity;
  dcterms:identifier "actual query string"^^xsd:string;
.
:code-prov:dataCubeToBeRefined
  a prov:Entity;
.
:code-prov:QueryWizard
  a prov:SoftwareAgent;
```
code-prov:retrieveRDF
  a prov:Activity;
  prov:used :linkedOpenDataEndpoint1, :linkedOpenDataEndpoint2, :mySparqlQuery;
  prov:wasAssociatedWith code-prov:QueryWizard;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:retrievedRDF
  a prov:Entity;
  prov:wasDerivedFrom :linkedOpenDataEndpoint1, :linkedOpenDataEndpoint2, :mySparqlQuery;
  prov:wasGeneratedBy code-prov:retrieveRDF;
.

code-prov:manipulateAndEnrichDataCube
  a prov:Activity;
  prov:used :retrievedRDF, code-prov:dataCubeToBeRefined;
  prov:wasAssociatedWith code-prov:queryWizard;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:manipulatedAndEnrichedDataCube
  a prov:Entity;
  prov:wasDerivedFrom code-prov:retrievedRDF, code-prov:dataCubeToBeRefined;
  prov:wasGeneratedBy code-prov:manipulateAndEnrichDataCube;
.

code-prov:DataExtractor
  a prov:SoftwareAgent;
  .

code-prov:transferEnrichedDataCubeToDataExtractor
  a prov:Activity;
  prov:used code-prov:manipulatedAndEnrichedDataCube;
  prov:wasAssociatedWith code-prov:queryWizard, code-prov:DataExtractor;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:annotatedTable
  a prov:Entity;
  prov:wasDerivedFrom code-prov:manipulatedAndEnrichedDataCube;
  prov:wasGeneratedBy code-prov:transferRDFToDataExtractor;
.
code-prov:conversionToDataCube
  a prov:Activity;
  prov:used code-prov:annotatedTable;
  prov:wasAssociatedWith code-prov:DataExtractor;
  prov:wasInfluencedBy :maxMuster;
.

code-prov:finalDataCube
  a prov:Entity;
  prov:wasDerivedFrom code-prov:annotatedTable;
  prov:wasGeneratedBy code-prov:conversionToDataCube;
.
Appendix E: Provenance Data for Scenario 5

```turtle
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix code-prov: <http://code-research.eu/ontology/provenance/> .
@prefix : <http://example.com/> .

:maxMuster
  a prov:Person;
  a code-prov:Consumer
  foaf:givenName "Max Muster"^^xsd:string;
  foaf:mbox <mailto:max@muster.org>;
  foaf:homepage <http://max.muster.org>;
.

code-prov:dataCubeToBeDisplayed
  a prov:Entity;
.

code-prov:visualizationMethod
  a prov:Entity;
  .

code-prov:Visualizer
  a prov:SoftwareAgent;
  .

code-prov:visualizeCube
  a prov:Activity;
  prov:used code-prov:dataCubeToBeDisplayed, code-prov:visualizationMethod;
  prov:wasAssociatedWith code-prov:Visualizer;
  prov:wasInfluencedBy :maxMuster;
  .
```