OPEN CITIES

cip competitiveness and innovation framework programme 2007–2013

DELIVERABLE

Project Acronym	**Open Cities**
Grant Agreement number:	**270896**
Project Title:	**OPEN INNOVATION Mechanism in Smart Cities**

D4.4.11 a

Definition of Data Sets & Scenarios

Revision: A, v1.6

Authors:

Edzard Höfig, Jens Klessmann, Nils Barnickel (Fraunhofer)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	**Public**	
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Organisation	Description
1.0	4.7.11	Ed	Fraunhofer	Initial Skeleton
1.1	4.7.11	Ed	Fraunhofer	Reformatting
1.2	11.7.11	Ed	Fraunhofer	Added input on Chapter 3 and 5, reformatting
1.3	15.7.11	Ed	Fraunhofer	Input on use cases
1.3b	16.7.11	Nils	Fraunhofer	Added generic scenario and input on use cases
1.3c	17.7.11	Jens	Fraunhofer	Added input on Chapter 2 and 3.3
1.4	18.7.11	Jens, Nils, Ed	Fraunhofer	Merged version
1.5	22.7.11	Jens, Nils, Ed	Fraunhofer	Formatted version
1.6	2.8.11	Jens	Fraunhofer	Included additional input by WP4 partners

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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.

TABLE OF CONTENTS

Synopsis

In WP4 of the Open Cities project, we promise to create a pan-European platform for management and publishing of Open Data. To get there, we need to have two things: We need data that we can publish and we need a software infrastructure to publish them. This deliverable details the progress that we made for these two areas. Regarding the collection of data, we describe the method that we used to collect existing data sources (a questionnaire based survey) and propose a selection of three data sets that can potentially be used beneficially with a pan-European platform. Regarding the engineering of the infrastructure, we designed a high-level scenario and a number of technical use cases based on the previously collected requirements (see D4.4.2). These use cases will guide us through the implementation phase for the Open Data platform functionality.

Abbreviations

API	Application Programming Interface
App	(mobile) Application
CSV	Comma-Separated Values
EU	European Union
HTML	HyperText Markup Language
IT	Information Technology
JSON	JavaScript Object Notation
OD	Open Data
OKF	Open Knowledge Foundation
PDF	Portable Document Format
RDF	Resource Description Framework
REST	Representational State Transfer
RSS	Really Simple Syndication
SPARQL	Simple Protocol and RDF Query Language
UPF	Universitat Pompeu Fabra
URL	Uniform Resource Locator
XLS	Microsoft Excel

1. INTRODUCTION

Creating an Open Data (OD) platform for the Open Cities project is a challenging task. We will need to arrive at a running, technical solution within a tight schedule (end of 2011) while ensuring that a sufficient amount of data assets will already be managed and provided through the platform. Thus, we have to do two things: On the one hand we have to identify the data sets that will be provided by the platform, on the other hand we will need to engineer the platform functionality itself.

For the first task we are using a questionnaire that helps us to survey existing data sources in the participating cities. From these sources we will choose a number of data sets that will be used for further research and within the pan-European challenges. We plan to also provide access to the other data, but we will not aspire to convert them to a common format or store them as part of the OD platform. They will only be searchable through the same interface, which ensures that they can be found and accessed in a homogenous manner.

For the second task we are using a high-level scenario to guide us through the engineering process. Based on the application of the scenario to a data set, we identify a number of roles that interact with the platform. We also create several technical use cases for each of the roles, which allow us to start with the engineering process by implementing each of the use cases.

2. THE OPEN DATA SURVEY

In order to identify data sources and different data sets within the partaking cities a survey was undertaken. As WP4 and WP6 have in some parts similar objectives, namely identifying data sets and providing access to them via certain IT platforms, the survey was conducted in close collaboration with WP6. For this the leaders of both work packages Universitat Pompeu Fabra (UPF) and Fraunhofer FOKUS aligned their survey development efforts.

2.1. METHODOLOGY

As a methodology for the OD survey a quantitative paper based questionnaire was chosen by the working team. The requirements for the correct approach were:

- Easy distribution among different stakeholders within each partaking city government.
- Remote conducting of the survey, as traveling to each partner and undertaking face to face interviews would have been too complex.
- Work packages 4 and 6 could integrate their surveys
- A large amount of standardized input was necessary. Gathering information about contacts and activities related to OD in each city within part one of the survey. Listing of available data sets in part two of the questionnaire.

As the main objective of the survey was to gather mostly information and not emotions or opinions on the topic of open government data a quantitative approach was chosen over qualitative phone or face to face interviews with different participants within in each city.
The questionnaire was developed and tested in close collaboration between researchers from UPF in Barcelona (WP6) and Fraunhofer FOKUS in Berlin (WP4). During the survey development phase feedback on initial versions concerning any further items, the wording or the order of questions was gathered from the partaking city administration representatives.

2.2. THE QUESTIONNAIRE

The questionnaire consists of two parts. In the first part the focus is on investigating the status quo of public sector information provision activities within each municipality. With the second part the amount and variety of local authority data sets should be determined.

Questionnaire Part One

The first part is structured into the four sub-categories:

- The Organization and Current Availability of City Government Data
- Ongoing Data Provision Activities
- City Government Plans
- Motivation & Technological Assistance

In the first subcategory the objective was to gather information about the different partaking city government organisations, already existing open government data offerings like central data platforms and whether responsibilities for OD exist within each organisation. The answers to these questions allow the Open Cities consortium to identify and contact relevant stakeholders in each city more easily.

In the second subcategory input about ongoing OD activities in the different cities was requested from the survey participants. Especially information about existing roadmaps or other strategic documents was of interest. This information helps the Open Cities consortium to better align with existing activities and join forces.

The goal of the questions in the third subcategory "City Government Plans" concerns the aspired local activites in connection with the Open Cities project. Questions were asked about the classes of information in which data will be provided during the project and the willingness to host different formats of data.

The fourth subcategory of Part I in the survey consists of questions about the reasons for pursuing OD, advantages in collaborating at an international scale on this topic and suggestions for support needed in order to achieve the provision of public sector information as planned.

Questionnaire Part Two

With the second part of the survey a finer grained look at data sets within each partaking city was aimed at. For this the city representatives were asked to provide information about individual data sources and data sets. A pre-defined spreadsheet was distributed among the city partners. This document was structured into data set topics and attributes for each data set. The objective of collecting more information about data sets within each city, was to identify data sets and have a better understanding of their types, amounts and formats. Naturally a comprehensive investigation of existing public sector information was neither feasible nor within the scope of the Open Cities project. The extracted information about data sets supports the requirement analysis of the planned OD platform and the later loading of the platform with data sets and information about them.

The OD spreadsheet consists of twenty-one worksheets. The first worksheet (labelled "General Info"), listed varying dataset attributes (see below for detailed description of each attribute), their corresponding definitions, and the varying classes of data (listed further below). A second worksheet (labeled "Examples") illustrated what sample responses may look like for three mock datasets. Furthermore, the spreadsheet contained an additional 18 worksheets, each of which corresponds to exactly one data class. Moreover, each of these worksheets contained pre-defined templates (composed of dataset attributes and corresponding cells to store individual responses). That is, one template for each dataset.

For each data set the following information was requested:

Dataset Attribute	Definition
Dataset Name	Enter an identifier for the dataset (e.g., filename, document name, etc.).
Data Manager Name and Contact Info	Enter the name and contact information (e.g., *phone number, email, …*) for the data manager (i.e., the primary person responsible for the data, who may or may not be the owner or creator of the data). This information will be needed in the event there are questions arising when reviewing the dataset.
Description	Write a **short description** that accurately reflects the contents in the dataset.

Data Formats	List the specific data formats that are supported. Possible answers include *html, csv, kml/kmz*, *pdf, shapefile, txt, xls, xml*.
Data Handling Rules	Describe the particular data handling rules/policies, if any, that must be followed. Possible answers include none or restricted (e.g., some special data handling policies apply). If restricted was selected, please specify the particular data handling conditions. Ideally, datasets will have a license, such as the Creative Commons *CC0 1.0 Universal* (http://creativecommons.org/publicdomain/zero/1.0/).
Listing of Factors	List the names of the factors in the dataset. For example, *year*, *square meter*, *time*.
Data Timeperiod	Indicate the *time period* for the data (e.g., data for 2009, data between 2005-2010, etc.).
Dataset Size	Indicate an estimate of the dataset size (e.g., less than 1 MB, 1 MB or greater, 1 GB or greater, etc.). If the exact size is known, please specify it (or in the case of a database, enter the number of records).
Data Access	Indicate whether one can gain access to the data via a **URL** (if known, please specify), an **API** (if relevant, please specify the API, possible answers include web-service SOAP, web-service REST), and/or a **database** (if relevant, please indicate the database product & version, possible answers include MySQL Workbench 5.2, Oracle Database 11g Release 2, ...).
Data Freshness	Indicate how often the dataset is updated. That is, does it get updated *hourly, daily, monthly, annually,* etc. If it is real-time data, please include the average data rate (e.g., expressed as 100 kbps, 1 Mbps).
Data Collection and Interpretation	Indicate the difficulty that may exist when collecting, and/or interpreting the data by users. Possible answers include easy, semi-challenging, and challenging. By easy, we mean data that is structured, expressed in a digital format, and available online via a URL. By semi-challenging, we mean data that is semi-structured, and expressed in a digital format. By challenging, we mean data that is unstructured, and/or unavailable in digital form.
Availability	Indicate whether the dataset already exists and is available. If not, indicate when it will be made available.
Language(s)	Indicate whether the dataset exists in only one language (or more than one language). Please specify the particular language(s).

Table 1: Requested data set attributes in part 2 of the Open Cities WP4 & WP6 OD Survey.

The participants were asked to fill in their responses according to 18 pre-defined data classes and one “Others” category.

Dataset Class	Examples
Arts and Recreation	Parks, Playgrounds, …
Business Enterprise, Economics, and Trade	Startups, Incubators, …
City Budget: Revenues & Expenditures	Distribution of Annual Budgets, Records of Annual Revenues and Expenditures, …
City Portal Web Statistics	Frequently Visited Webpages, …
Construction, Housing, and Public Works	Ongoing Projects, Planned/Future Projects, …
Crime and Community	Crime Statistics, Safety Initiatives, …

Safety	
Demographics	Registry Services, Population, …
Education	Elementary, Secondary, Higher Education (Universities), Libraries Wifi...
Elections	Polling Stations, Election Results, …
Emergency Services	Police, Fire, Medical, …
Energy and Utilities	Energy Demand, Water Consumption, …
Environment, Geography and Meteorological	Mapping and Geospatial Data, Air Quality Indexes, Water Quality Test Results, Weather, Climate Studies, …
Health and Disability	Annual Healthcare Costs, …
Labor Force and Employment Market	Size of the Labor Force by Industry, …
Law Enforcement, Courts, and Prisons	Citations Issued Annually, Number of Court Cases Annually, …
Political	Zone Redistricting, Record of Decisions Made by Politicians in the Current Year, …
Tourism	Annual Number of Visitors, Frequently Visited Museums, …
Urban Transport	Transportation Schedules, Bike Paths, Bike rental service …
Others	…

Table 2: Overview of the pre-defined data classes used within the Open Cities project WP4.

2.3. SURVEY RESULTS

The results from the conducted survey are described in this chapter. The questionnaire was distributed via e-mail to the five partaking cities of Amsterdam, Barcelona, Berlin, Helsinki and Paris on 4. January 2011 with a return deadline of January 31, 2011. Five respondents filled out part 1 of the survey. Four respondents also filled out part 2 (compare Table 3). As not all cities responded to every part of the survey and the listed data sets do not necessarily represent all available data sets on one topic the following results can only be seen as indicative.

City	Part 1	Part 2
Amsterdam	Completed	Completed
Barcelona	Completed	Completed
Berlin	Completed	Completed
Helsinki	Completed	-
Paris	Completed	Completed

Table 3: Overview of return results of the Open Cities WP4 & WP6 OD Survey

Paris named two open data officers, Barcelona one (Table 4). The others of the responding cities had not appointed an open data officer at the time of the survey. The responding cities identified certain individuals as proponents of public sector information provision though. In the survey it was asked, whether Open Data Evangelists could be named (Table 5).

City	Open Data Officers
Barcelona	Lluís Sanz (City of Barcelona, Institut Municipal d'Informàtica)
Paris	Georges-Etienne FAURE(Mayor's Deputy Adviser for innovation, Universities and Research); Jean-Philippe CLEMENT (Commissioner for IT, General-Directorate)

Table 4: Named open data officers in the responding cities

City	Open Data Evangelist
Amsterdam	Katalin Gallyas (Economic Affairs Amsterdam EZ), Frank Kresin (Waag Society)
Barcelona	Isaac Aparicio (City of Barcelona)
Berlin	Daniel Dietrich and Friedrich Lindenberg (Open Data Network Germany)
Helsinki	Ville Meloni (Forum Virium Helsinki)
Paris	Jean-Louis MISSIKA (Deputy Mayor for innovation, Universities and Research)

Table 5: Named open data evangelists in the partaking cities

In part I.B of the survey information about ongoing data provision activities was gathered. The cities are at different stages concerning the implementation of strategies, procedures and technical solutions for provision public sector information according to the open data principles.

The City of Amsterdam described two ongoing or planned open data activities in their response. First a pilot activity for transforming local statistical data into linked data. For this pilot the Department for Research and Statistics of the City of Amsterdam (O+S) and the Free University of Amsterdam cooperated. The second activity was the then planned launch of www.apps4Amsterdam.nl.

In the response from the City of Barcelona it is stated, that the city plans establishing an Open Data Commission as part of the Barcelona City Council. This commission will be responsible for developing actions related to the opening of data by Barcelona City Council.

Berlin described in its response two future activities. One is a conceptual-strategic study, looking at the Berlin situation concerning open government data in detail. The second mentioned activity is the preparation for open data testing projects in cooperation with different stakeholders in the city.

The response from Paris delineates that the City of Paris has launched a platform for publication of the first datasets. The City also plans to organize a barcamp around its approach and its Open Data datasets. The City will organize a competition to develop web services and mobile applications using their datasets and to provide a new service to its users.

The main focus of the cities for rolling out data in the context of the Open Cities project is on statistical data about the demographic population development and electoral data (Figure 1). Answers given by three of the cities were data on the city budget and business enterprise, economics, and trade data.

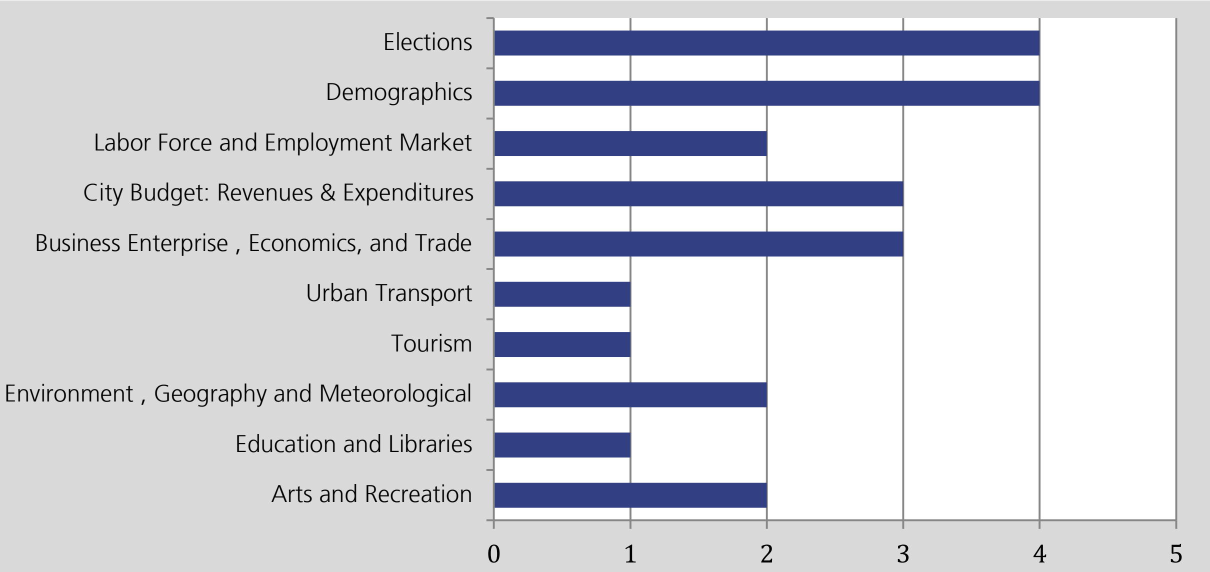

Figure 1: Answers to question 7. Indication of planned classes of data to be rolled out for the Open Cities project.

The partaking cities want to host services for raw data access, but they do not want to host services for linked data access (Questions 8 and 9). This could be due to the fact, that linked data is a relatively novel concept for most administrative organisations.

The partaking cities see mainly economic stimuli and greater transparency of public sector processes and decisions as reasons for pursuing open data (Figure 2). This seems to coincide with the most often mentioned possible advantages of the open government data approach. The third, often mentioned, reason for following the open data path, reaching greater organisational efficiency is only cited by two local authorities as relevant.

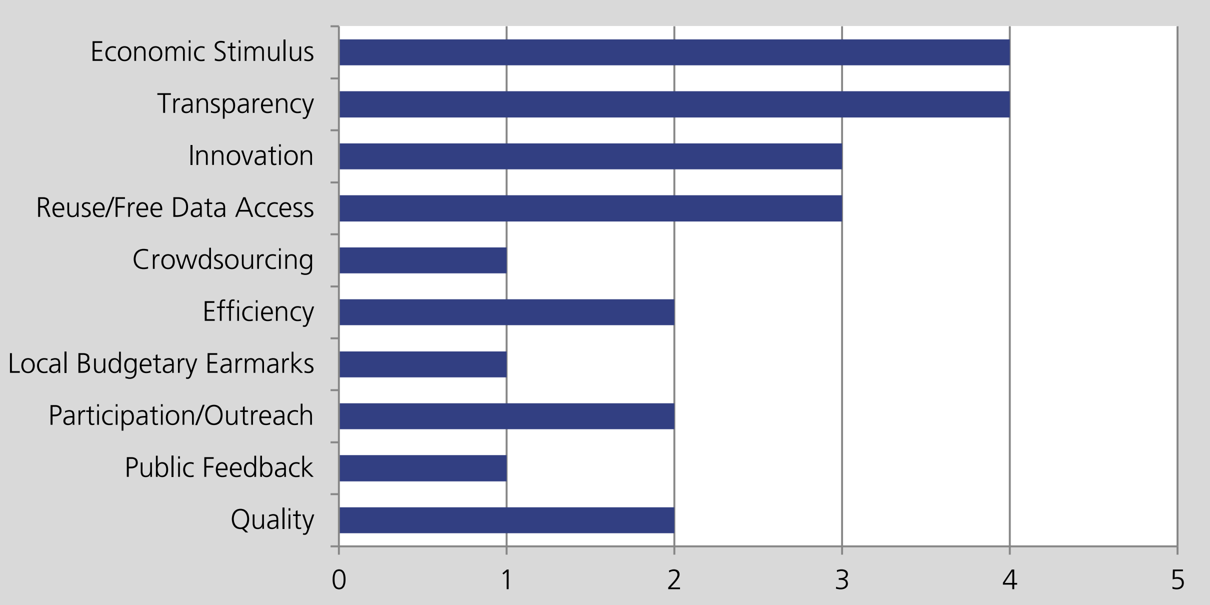

Figure 2: Answers to question 10. Indication of the respondents reasons for pursuing open data.

Asked, what advantages each city sees in collaborating internationally in the field of Open Data, the respondents expect especially harmonized licensing rules (Question 11). This would help authorities in deciding upon the right license to use for their data sets and support legal interoperability of the data released as open data. As further expectations the respondents named exchange of knowledge with other experts, identification of best practices and feedback on their own activities.

The cities seek mainly technological support in deciding upon the right platform to use and implementing such a platform (Questions 12). The cities also mentioned input on best practices for pursuing the open data approach, support with spreading the Open Data initiative and the possibility of having the data in the cloud.

In part II of the Open Cities survey the cities were asked to identify individual data sets in order to gain a more fine grained comprehension of the current "data situation" in each city. In the following a short overview across all mentioned data sets will be given.

Looking at the data sets, which the cities reported per data set topic (Table 6), the top five areas are law enforcement, demographics, environment & geography, education andurban transport. The least data sets per data set topic can be found in emergency service, city portal statistics, crime & community safety and arts & recreation.

	Amsterdam	Barcelona	Berlin	Paris	Total
Arts & Recreation	1	-	3	-	**4**
Business Enterprise	3	4	19	-	**26**
City Budget	1	-	8	-	**9**
City Portal Statistics	-	-	-	2	**2**
Construction & Housing	2	9	5	1	**17**
Crime & Community Safety	2	-	1	-	**3**
Demographics	1	53	12	-	**66**
Education	2	3	27	1	**33**
Elections	1	8	6	1	**16**
Emergency Services	1	-	-	-	**1**
Energy & Utilities	1	-	10	1	**12**
Environment & Geography	4	45	13	1	**63**
Health & Disability	2	-	12	-	**14**
Labor Force & Employment	3	-	5	-	**8**
Law Enforcement	-	167	3	-	**170**
Political	1	-	4	-	**5**
Tourism	2	-	2	-	**4**
Urban Transport	2	13	12	-	**27**
Others	6	73	4	1	**84**
Total	**35**	**375**	**146**	**8**	

Table 6: Data sets per data set topic as reported by the responding cities

From the reported data sets most sets are in the native local language (Table 7). Amsterdam and Berlin provide a few data sets in English, but the majority is in either Dutch or German. Barcelona provides its mentioned data sets in Catalan. The RDF resources are reported as being in English though. Paris reported no data sets in English as did Helsinki.

	Amsterdam	Barcelona	Berlin	Paris
Catalan	-	375	-	-
Dutch	35	-	-	-
English	12	-	5	-
French	-	-	-	8
German	-	-	146	-

**Table 7: Data sets per language as reported by the responding cities.
Barcelona: RDF resources in English**

Concerning the data set formats in each participating city, most data sets are provided in Microsoft Excel (XLS), Hypertext Markup Language (HTML) or Portable Document Format (PDF). The next largest form of representation of data sets is via databases. Semantic web formats like the Resource Description Framework (RDF) were only mentioned by Barcelona.

	Amsterdam	Barcelona	Berlin	Paris
CSV	11	16	-	4
XLS	18	85	60	2
PDF	-	240	83	-
XML	1	11	-	-
KML/SHAPE	2	-	1	2
DATABASE	-	-	28	-
HTML	-	-	88	-
ODF	-	-	-	3
RDF	-	12	-	-
ZIP	-	20	-	-
TXT	-	1	-	1
Others	1 (ASCII)	1 (BIN)	-	-

Table 8: Data sets per data set type as reported by the responding cities

3. SELECTED PAN-EUROPEAN DATA SETS

Our ultimate goal is to release all public data sets from different cities across Europe in a compatible format. Unfortunately, such an effort will take a far longer time than possible within the limits of the Open Cities projects. Thus, we decided to begin with only a very limited number of data sets (one to three sets) and to exercise the complete process for data collection, data format homogenisation and data publishing on only these. This decision does not impact the general idea of opening all possible data sets, as we still strive to create a supporting infrastructure for publishing data sets in arbitrary formats.

We propose to start with three data sets: our primary idea is to use information about tourist sites (points of interests) as such data is readily available, does not have privacy issues (e.g. personal references) and there is also an interest for publishing these data sets on part of the cities. There are two secondary data sets that we would like to use: Information on the city budget and spending and demographic data. Spending data is an area with a specific interest for the general public and there are currently efforts within the Open Knowledge Foundation (OKFN) to make more of this data available. The demographic data was chosen due to the ease of collection (such data sets are usually published openly) and the promise of a good compatibility between the data sets.

3.1. PRIMARY EXAMPLE: TOURIST SITES

Data sets of this kind will contain at least 50 of the touristic attractions per contributing city, according to a single format.

Motivation

As we will use the data sets as part of the pan-European challenges conducted in Task 4.5, having data sets that are of interest for a geographically dispersed audience is of importance. With data about tourist attractions, we believe to have found such a data set: for example, visitors from Barcelona will have an interest in information about tourist attractions when coming to Berlin and the other way round. Having data sets that are relevant for a wider European audience is an important economical factor for mobile application development. The overall effort for developing an app is more justified when developing for a larger market, such as the EU, as when developing an app for only a single city.

Privacy is perceived as a big issue in regard to publishing OD. By concentrating solely on tourist attractions, we believe to not introduce any problems related to personal references or the like. The information contained in these data sets is generally publically available and not sensitive to security concerns. There is a strong will to make this data public, as tourism will directly benefit from the availability of this information.

A problem in connection with publishing such data is the administrative organisation of the responsibilities for such data sets. A common construct for a city is to employ an independent organisation with touristic marketing. As with other data sets (e.g. transport data), such independent organisation might be hard to convince to contribute to openly contribute their information, as their business models are usually based on earning money through sale of the data. As part of the Open Cities project, we do not believe this to be a problem. We are planning to release only quite rudimentary information on well-known landmarks, with no additional information that could represent an “added value” and therefore be considered a challenge to companies for touristic marketing.

Data Set Structure

In order to achieve interoperable data sets for touristic attractions from each city a common metadata set structure has to be established. Currently information concerning tourist attractions is structured and presented differently in each city. By conducting a comparison between different data sources describing such tourist attractions a preliminary structure for their description was derived. The evaluated data sources were corresponding websites in the cities of Berlin, Paris and Amsterdam (Table 9). Each city here lists its top attractions and provides basic information about the relevance, geographical location, and access by different modes of transport, opening hours or entrance fees. For further comparison the metadata structure of analogous articles from Wikipedia were analysed.

Evaluated data source	URL
Amsterdam Places to go	http://www.iamsterdam.com/en/visiting/placestogo
Amsterdam Venues	http://www.iamsterdam.com/en/whats-on/venues/
Berlin tourist attractions	http://www.berlin.de/orte/sehenswuerdigkeiten/
Paris Museums & Monuments	http://en.parisinfo.com/museums-monuments-paris/

Table 9: Evaluated data sources

The result of the comparison is a metadata structure qualified for describing tourist attractions in the different cities in a similar manner (Table 10). 17 fields were identified, of which four are seen as the minimal necessary information and 13 as optional information.

Title: In this field the name of the tourist attraction is given, for example: Eiffel Tower or Brandenburg Gate.

Geocoordinates: The geographic location of the attraction is specified by values (latitude, longitude) from the geographic coordinate system. The geographic location can be derived from the objects physical address.

Address: The physical address of a tourist attraction like: Am Kupfergraben 5, 10117 Berlin, Germany.

District: The part of the city in which a tourist attraction is located, like Paris Montmartre.

Description: A free text delineation of the attraction in question. The objective would be to include information about the cultural relevance of the site, its history or architectural style. Ultimately the information in this field will vary to a great extent, as the cities deem different aspects as important.

Other Information: A free text description focussing more on practical details surrounding a visit of the site by tourists. This could include information about available guided tours and merchandising products, etc.

Public transport: Description how to best reach a tourist attraction by public transport including mentioning of the closest stops for busses, subways or light rail.

Close by: In order to support the planning of a visitation route, information about other tourist attractions in close proximity to the current site can be given.

URL_entry: An HTTP URL referencing a resource with further information about the attraction site. This could be website within a general tourist information portal maintained by the city.

URL_map: An HTTP URL referencing a map displaying the tourist attraction in question.

Telephone: A phone number within the city, where additional information concerning the attraction can be obtained. Ideally this would be a support number where questions in foreign languages can be answered.

E-Mail: A general contact e-mail, where answers to questions concerning the tourist attraction can be given.

URL_site: An HTTP URL referencing a resource maintained by the responsible organization of the site. Oftentimes tourist attractions have their own website, like the Berlin TV tower.

URL picture section: An HTTP URL referencing a picture or a section with picture of the tourist site.

Opening hours: Current information about the opening and closing times of a tourist attraction are important information for foreigners. In addition exceptions from the opening rules can be mentioned.

Accessibility: Guests with varying types and degrees of disabilities can use information about how accessible the tourist attraction is in advance, in order to plan their visit.

Entrance fees: Price information is an important aspect of visiting a tourist attraction. Here the pricing scheme of a tourist attraction can be described.

Field name	Bindingness
Title	obligatory
Geocoordinates	optional
Address	obligatory
District	optional
Description	obligatory
Other Information	optional
Public transport	optional
Close by	optional
URL_entry	obligatory
URL_map	optional
Telephone	optional
E-Mail	optional
URL_site	optional
URL picture section	optional
Opening hours	optional
Accessibility disabled persons	optional
Entrance fees	optional

Table 10: Proposed metadata structure for tourism data sets

3.2. SPENDING DATA

The OKFN proposed to use public spending data as common data sets in a recent phone conversation[1]. This can be beneficial due to the current interest in the topic and might also be a good choice in regard to the availability of data. For some of the cities, data of this kind is already available (London, Berlin, Helsinki, among others) and there are a number of projects aiming at a user-friendly visualisation of the data sets, both for national budgets[2] and city budgets[3].

The general structure of data sets that cover spending data is simply: They contain a list of expenses, attributed to categories, purposes and/or administrative institutions. The entries can usually be structured hierarchically, e.g. an entry does not only belong to a single certain category or purpose, but each category is subdivided in a number of sub-categories. For example, a category that summarises environment costs could have a sub-category for money spent on environmental protection. Apart from the categorisation, it would be beneficial if such data sets would have explanatory remarks attached for each entry, or at least, for each (sub-)category. Otherwise, the data will only be meaningful to a small number of experts, but not the general public. Additionally, it would be good to have data not only from a single year, but from a number of previous years. This

[1] Project-wide phone conference on 11. July 2011
[2] For example: http://bund.offenerhaushalt.de, http://www.ukpublicspending.co.uk, http://www.openspending.org, http://wheredoesmymoneygo.org/
[3] For example: http://berlin.offenerhaushalt.de/dataset/berlin,

would enable an analysis of the changes of budget allocation over the years and promises the creation of interesting applications.

From the conducted survey in the Open Cities project, we know that at least Berlin and Helsinki plan to release city budget data. If we decide on using this kind of data set, all the participating cities have to agree on a common set of categories and a common method for categorisation used for structuring the spending data. Otherwise, there will be no comparability between the data, which makes creating a common, pan-European App based on these data sets difficult.

3.3. DEMOGRAPHICS

The demographic data was chosen due to the ease of collection (such data sets are usually published openly) and the promise of a good compatibility between the data sets. Each partaking city already provides data on its demographics via a local organisation concerned with population statistics. At the same time demographic data is to some extent standardised, which makes it easier to compare across different jurisdictions on the local and cross-national level.

Crucial for using demographic data as the basis for the development of interesting web or mobile applications is a high granularity of the data. Having for example demographic data only at the national level, will make it more difficult for application developers to conceive exciting solutions. At the same time the degree of granularity will have strict limits, as privacy aspects play a pivotal role in this field.

From the Open Cities survey it can be derived, that all partaking cities are either already providing demographic data or are planning to do so within the project timeframe. If we decide on using this kind of data set, all the participating cities have to agree on a common set of categories and a common method for categorisation used for structuring the demographic data.

4. HIGH-LEVEL SCENARIOS

4.1. GENERIC SCENARIO

In the following section we try to abstract from the specific scenario of tourist sites data, spending data and demographics data. The goal is to derive a generic scenario for OD provisioning based on the to be developed Open Cities platform. Figure 4 below highlights the major steps from a process-oriented perspective in this generic scenario and identifies the involved key actors and domains as well as key IT systems required.

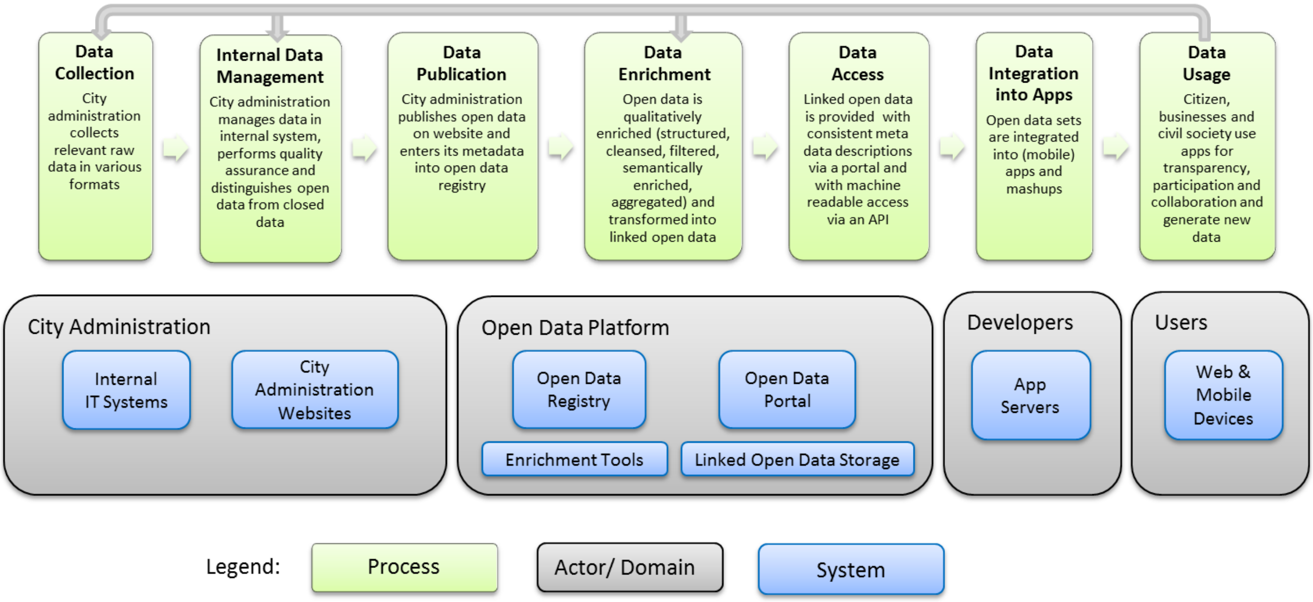
Data Collection
City administration collects relevant raw data in various formats
Internal Data Management
City administration manages data in internal system, performs quality assurance and distinguishes open data from closed data
Data Publication
City administration publishes open data on website and enters its metadata into open data registry
Data Enrichment
Open data is qualitatively enriched (structured, cleansed, filtered, semantically enriched, aggregated) and transformed into linked open data
Data Access
Linked open data is provided with consistent meta data descriptions via a portal and with machine readable access via an API
Data Integration into Apps
Open data sets are integrated into (mobile) apps and mashups
Data Usage
Citizen, businesses and civil society use apps for transparency, participation and collaboration and generate new data
City Administration
Internal IT Systems
City Administration Websites
Open Data Platform
Open Data Registry
Open Data Portal
Enrichment Tools
Linked Open Data Storage
Developers
App Servers
Users
Web & Mobile Devices
Legend:
Process
Actor/ Domain
System

Figure 3 Generic Scenario for Open Data Provision

The first step covers the initial *Data Collection*. Civil servants collect relevant raw data in various formats using sector specific procedures, which are not further analysed within this scope.

The second step deals with the *Internal Data Management* to highlight that the origins of OD are mostly[4] city administration internal IT systems. Therein, civil servants need to perform quality assurance and prepare data for further internal or non-OD related tasks and duties. Raw data which may contain, e.g. personalized information are made anonymous and aggregated to statistical data sets. Finally, the city administration has to distinguish between the data that remains internal and that data which can be published as OD.

The following step then describes the *Data Publication*. The respective departments of the city administration publish the selected OD sets on their websites, which may remain decentralized. However, in order to provide a single point of reference enabling consistent search and navigation across multiple OD sources, certain descriptions of the published data sets – the metadata – are registered in a city-wide OD registry.

[4] User or citizen generated open data are discussed later on including feedback mechanisms related to published data

The next step, *Data Enrichment*, constitutes an optional activity which increases the usability and thus the potential for OD to be effectively used in mobile apps and Web-based mashups. To enable machine-readability and enable to process and interpret OD seamlessly, data sets can be transformed into semantically richer formats such as linked data, which explicitly express the context and the relation between data. Using linked OD and corresponding technologies allows for sophisticated filtering, cleansing or structuring as well as combining and mapping different vocabularies used in different city departments to derive aggregated data sets. Such linked OD sets can be stored centrally on the OD platform, which provides the required tools and storage services. Thus, semantics-based data enrichment provides the foundation to enable apps to perform data fusion and generate new insights into OD not visible from an isolated perspective.

Accordingly, the next step describing *Data Access* not only mentions the channel of a Web-based portal enabling a consistent (based on standardized metadata) access to OD. Furthermore, the machine-readable access via an Application Programming Interface (API) is essential.

Then *Data Integration into Apps* can be performed by developers from private sector or civil society who build (mobile) apps and mashups and distribute them over external infrastructures including app stores and app servers not part of the city's OD platform.

Finally, *Data Usage* closes the lifecycle of OD. On the one hand citizens, businesses and civil society use apps via Web and mobile devices for transparency, participation and collaboration. On the other hand they also generate new data, which either can be considered in the initial *Data Collection* step or during *Internal Data Management* and quality assurance and as well during *Data Enrichment*.

The process-oriented approach and the first identification of actors, domains and key IT systems in one big picture provides a general understanding and overview. Based on this perspective the more detailed requirements analysis for the Open Cities data platform can be performed in the following section.

5. PLATFORM USE CASES

During the design phase for the OD platform, we derived a number of use cases from the requirements document [1]. The use cases have been grouped into five sets, each one relating to a different actor role. Consequently, there are currently five roles in the system design.

5.1. MAINTAIN DATA ASSETS

The actor for maintaining data assets is called the data steward. His task is to transform existing OD sets registered in the OD platform into linked data to provide seamless machine-readable access for Web mashups and mobile apps. The use cases for maintaining data assets are shown in Figure 4.

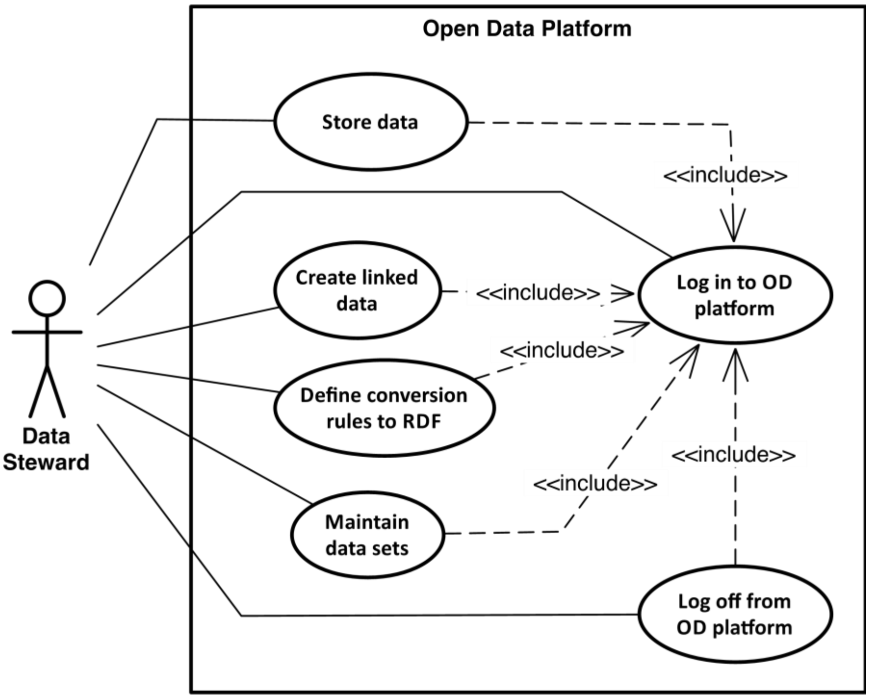

Figure 4) Maintaining Data Assets

Log in to OD platform

The data steward logs in to the OD platform with his user credentials (login account name and password) to ensure controlled access and data protection.

Create linked data

To enable seamless machine-readable reuse of OD sets for Web mashups and mobile apps selected frequently used data sets are transferred into linked data. The data steward gets the relevant data sets from the original sources using the metadata and references in the OD registry. By applying a predefined set of tools described in terms of a user guide from the Open Cities project the data steward transforms originally published OD sets in formats such as Excel, CSV or from a relational database into linked data. Furthermore, the resulting data sets are semantically linked to other data sets to enable contextual data aggregation and processing by app developers. As the transformation is a recurring effort for OD with the same format and structure, the data steward uses RDF conversion rules, which allow to semi-automate the transformation from original data sets to RDF-based linked data.

Define conversion rules to RDF

To support the transformation process of creating linked data, the data steward defines conversion rules, which semi-automatically support the transformation from data sets in less machine-readable formats to RDF-based linked data that can be seamlessly reused in application development. Store DataThe result of the transformation process are copies of the original data sets expressed in the RDF-based linked data format. The resulting linked data are stored in a triple store and are made accessible via an API.

Store Data

The converted linked data are stored in a persistence layer within the OD platform.

Maintain data sets

The data steward ensures the quality of the stored linked data along its life-cycle including updating of data amendments and corrections. This task includes as well the maintenance of links between linked data to ensure that they stay accurate and up to date.

Log off from OD platform

The data steward logs off from to the OD platform to ensure that no uncontrolled access is provided to the OD platform.

5.2. MAINTAIN METADATA

The responsible actor for maintenance of metadata is the data owner. His task is to maintain the life-cycle of metadata entries describing the data sources under his authority.The use cases for maintaining metadata are shown in Figure 5Figure 4.

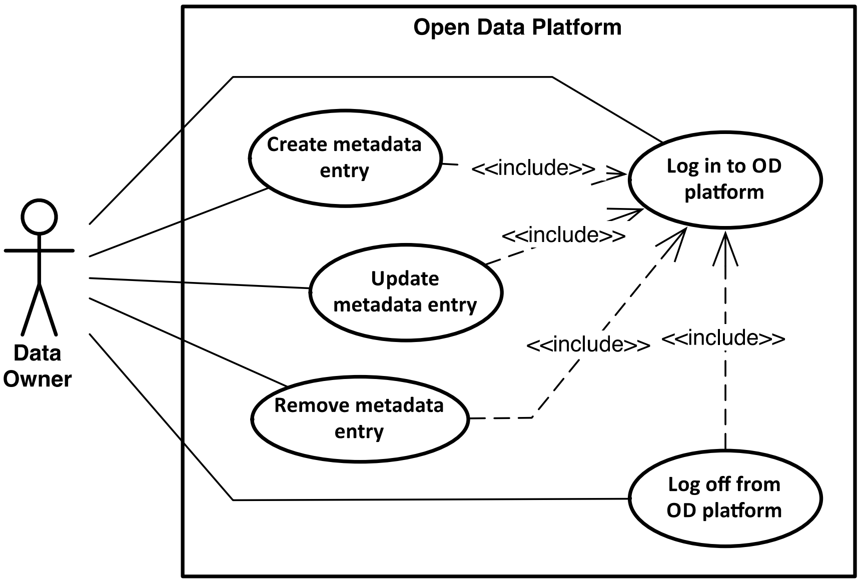

Figure 5) Maintaining Metadata

Log in to OD platform

The data steward logs in to the OD platform with his user credentials (login account name and password) to ensure controlled access and metadata protection.

Create metadata entry

The data owner registers a decentralized published OD set (e.g. on a City's administration website) in the OD platform. Therefore, he uses the Open Cities metadata standard to describe the OD set including e.g. the data source, the categories it belongs to or the references where the data sets can be downloaded and the respective format. The metadata entry is stored within the OD platform in a registry.

Update metadata entry

Once the metadata changes, e.g. due to the provision of additional data entries for an existing data set, the data owner updates the metadata entry and saves the changes in the OD platform's registry.

Remove metadata entry

In order to cover the whole life-cycle of OD provisioning, the data owner also needs to remove metadata entries once the related OD set is removed from the city's administration website or from the data storage of the OD platform.

Log off from OD platform

The data owner logs off from to the OD platform to ensure that no uncontrolled access is provided to the OD platform.

5.3. PLATFORM ADMINISTRATION

The use cases for administration of the platform are shown in Figure 6. The role "Platform Administrator" is responsible for configuration and maintenance of the OD platform, but not for the data assets published through the platform. An IT technician would usually cover this role.

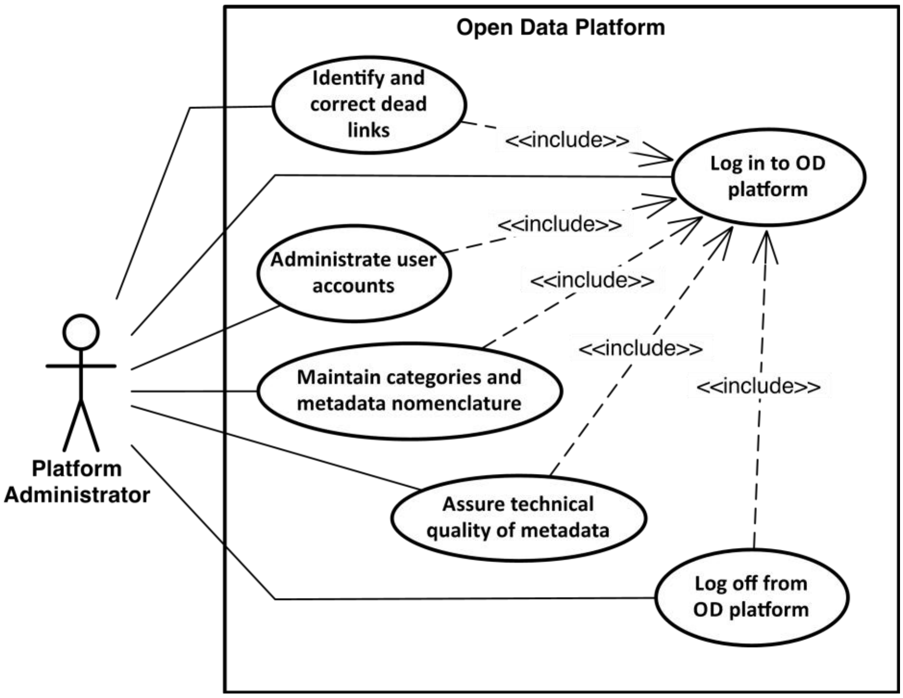

Figure 6) Administration of the OD platform

Log in to OD platform

This use case is similar to the use case of the same name in Section 5.1

Identify and correct dead links

During operation of the platform, a number of links are created. Links are particularly uses for referencing data sources from their metadata entries, but they can also appear in comment fields, data asset descriptions, and they are used to identify Apps for a certain data set. It is in the nature of Uniform Resource Locators (URL) that they are only maintained unidirectional, meaning that they can break once the referenced location changes. As this is a major problem, we are proposing to use a continuous, automatic process that discovers broken links and informs the Platform Administrator, who can then either correct the link herself (e.g. in the case of a changed server name or data set location) or notify the responsible Data Steward (e.g. when knowledge about the data set is needed to reconstruct the link).

Administrate user accounts

Access to the platform functionality is regulated through authorisation processes that are based on registered user accounts. While it is not necessary for Platform and Application Users (see Sections 0 and 5.5) to conduct an identification step when accessing the platform, it is required for roles that are able to modify the published data assets (Data Owner and Data Stewart). The Log in to OD platform use case is used to establish the roles of users. Due to security implications, the affiliations of people with these roles should be controlled. This is done through management of user accounts, which are created and revoked in accordance to given management processes for the platform.

Maintain categories and metadata nomenclature

Any metadata that is stored in relation to data assets needs to conform to a given nomenclature. The reason behind this is to assure the mutual compatibility between stored data assets, at least in regard to searchability and management of the data sets. The same argument is true for the set of categories that data assets can be assigned to. Both, categories and metadata, needs to be centrally maintained in accordance to management processes for the platform.

Assure technical quality of metadata

As metadata entries are created manually by, potentially, a large number of people, a certain quality standard should be enforced. This includes the automatic checking of metadata entries for, e.g., mandatory fields and field formatting, but also the marking of exiting metadata sets for correction in case of low-quality metadata.

Log off from OD platform

This use case is similar to the use case of the same name in Section 5.1

5.4. PLATFORM USE

The use of the platform is shown in Figure 7. Users that conform to the Platform User role are not required to identify themselves, as the read-only access to all published data assets should be possible for everyone.

Search / Browse metadata

The metadata catalogue that is maintained by the OD platform has to be searchable. We envision two different methods for search access: a “Google-style” search box with an optional extension for advanced search parameters (e.g. restriction to keyword or title) and a catalogue interface that allows Platform Users to browse the metadata (e.g. by navigating though a category hierarchy).

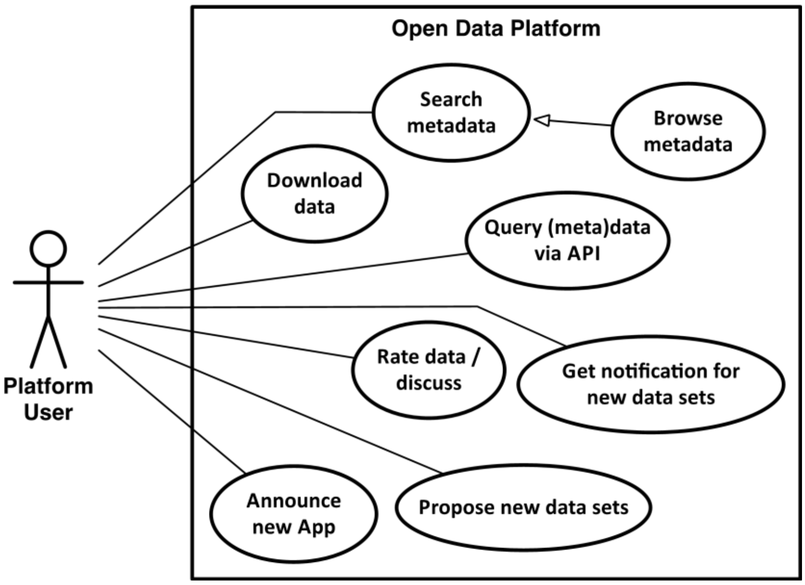

Figure 7) Usage of the OD platform

Download data

Once a Platform User has identified a data asset through the Search metadata use case, the Platform user can proceeded to access the data linked by the metadata entry. We do not specify how to access the data, but imagine that the usual mode of access would be through a clickable URL that directly links to a document containing the data.

Query (meta)data via API

It should be possible to also access the metadata catalogue programmatically by means of an API. We plan to rely on web technology for this and propose to expose the catalogue both by means of a REST based interface using JSON, as well as a SPARQL endpoint exposing an RDF representation of the metadata.

Get notification for new data sets

The Platform User needs to be informed of the introduction of new data sets, respectively changes or updates to existing data sets. The platform will expose RSS / Atom feeds for this purpose, which can be accessed with suitable newsreaders or aggregators.

Rate data / discuss

It will be possible for a Platform User to comment on a given data asset and to give feedback on the data catalogue. There are various instruments that can be used for this purpose: We can allow for user commenting on each data set or provide a separate discussion forum. The platform will also expose information that will make it possible to get in touch with the responsible Data Stewart for a given data set.

Propose new data sets

There will be a feature that allows contributing feedback regarding missing information in the data catalogue. Platform Users should be enabled to submit proposals for new – not yet openly published – data sets.

Announce new App

Platform Users will be able to submit information about apps that work with the information contained in certain data sets. The idea is to provide users with links to suitable applications, directly from the web pages that describe a data set and vice versa.

5.5. APPLICATION USE

The use of applications is shown in Figure 8. The Application User role covers users that interact with the OD platform indirectly, using third-party Apps. A part of the platform will be designed as an application portal, where Application Users can inform themselves about applications that use the data provided by the OD platform. The apps are not hosted on the portal itself, but can be accessed from dedicated app stores by means of links.

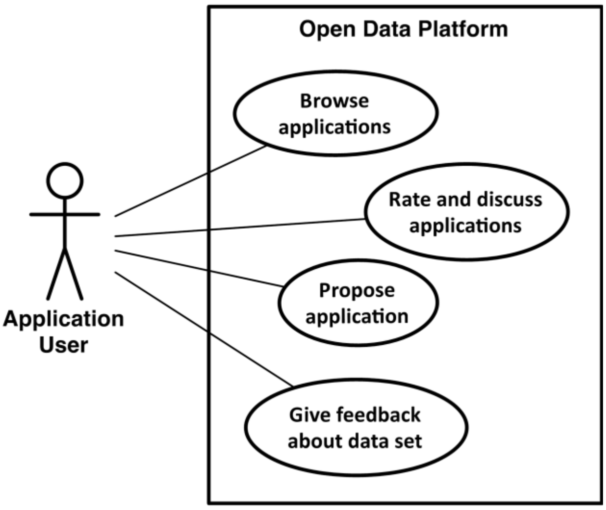

Figure 8) Use of applications

Browse applications

The platform hosts a structured listing of apps, which can be browsed by app category. The Application User can see details to each app and may choose to download it using a provided link to an app store. Furthermore, the data set that an app deals with are displayed in the app information web page, as well

Rate and discuss applications

To enable other users a better orientation regarding the quality of the apps, we plan to create a mechanism that allows Application Users to rate applications by means of a five star scale. There will also be an option to discuss an app using individual comment threads, or possibly a separate forum.

Propose application

Application users will have a dedicated place, where they can give feedback on apps that they would like to see being developed. This information can be used by all of the OD platform roles to optimise the offerings, both in regard to data sets, as well as to apps.

Give feedback about data set

During the usage of an application, users may encounter errors in the underlying data sets. We will provide a way for Application Users to comment on the data sets, so that data set quality can be improved.

6. CONCLUSION

The past chapters gave an overview of the status of the Open Cities project in regard to the identification and definition of data sets, as well as to high-level operational scenarios of the planned OD platform.

To identify the existing public data sets of the partner cities, we first conceived a questionnaire and executed a survey. These activities were lead by Fraunhofer FOKUS and UPF, with contributions coming from the cities of Amsterdam, Berlin, Helsinki (Forum Virium) and Paris (Cap Digital). Based on the received feedback, we presented an overview of a number of data sets, which could be used as OD within the Open Cities project.

In the next step, we propose three different data sets, one as a primary example (tourist sites) and two secondary examples (spending data and demographics) that can be used to test the European federation aspects of the OD platform. The idea is to capture these datasets in a RDF format prescribed by a common ontology and we also want to use these to driven the developments of Apps as part of the pan-European challenge organised in task 4.5.

Deciding on the structure and content of the data sets to use is only a part of our preparation work for the platform. The second part consists of identifying the dynamic aspects that the data is subjected to. We presented a high-level scenario based on the primary tourist site example and subsequently deducted a more generic scenario. This serves as a tool for setting the stage for the subsequent chapter that documents the platform functionality using, more technical, specifications of a number of use cases for the OD platform.

We propose to use five different roles (actors) that deal with the platform: the Data Stewart, Data Owner, Platform Administrator, Platform User and Application User. For each role, use cases are described. Based on these descriptions we will now develop the concrete technical functionality that is provided by the platform.

In summary: Based on the feedback that we got from conducting the survey, we were able to collect a larger number of available data sets. We were also able to identify the interests of the project partners in regard to future publication of data. We are now proposing three data sets that we can use as guidelines in a pan-European scenario. In respect to the platform functionality, we created a high-level scenario and derived a number of concrete use cases, which have already been presented at the last consortium meeting (in Paris, June 2011).