"Linked Open Apps Ecosystem to open up innovation in smart cities"

Project Number: 297363

Deliverable:	**D 4.4 Open Data**
Version:	**V1.1**
Delivery date:	**30/01/2013**
Dissemination level:	**PU**
Author:	Yury Glikman, Linda Strick, MichaelStemmer,Florian Weigand (**FRAUNHOFER)**
Reviewer:	**UOC, CITILAB**

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

Summary

This deliverable provides an overview of Fraunhofer Open Data Platform and Open Data platforms of London, Barcelona, Genoa and Bologna, and presents an approach to their integration in the iCity platform on the basis of Fraunhofer Open Data Platform.

DOCUMENT HISTORY

Version	Date of issue	Status	Content and changes	Modified by
V0.1	05/10/2012	Draft	Creation, ToC	Fraunhofer
V0.2	14/11/2012	Draft	Overview of OD platforms added	Fraunhofer
V0.3	20/12/2012	Draft	Integration concept is defined	Fraunhofer
V1.0	25/01/2013	Final	Document editing	Fraunhofer
V1.1	30/01/2013	Final	Update to address the comments of peer-reviewers	Fraunhofer

TABLE OF CONTENTS

Abbreviations and Acronyms

Acronym	Description
OD	Open Data
RDF	Resource Description Framework (RDF), www.w3.org/RDF/
SPARQL	Protocol and RDF Query Language

1. Introduction

Making public data accessible to the citizens has been an important topic for public administrations of different levels for many years. There have been many efforts on the regional and national level and many initiatives were supported by the European Commission. At this stage, many European cities already have their own Open Data platform and many are currently developing one. Some, the most advanced of them, understand exactly what they need and are already considering moving to a platform of a new generation, which could provide them with the desired functionalities. However, most of the cities still do not have their own Open Data platform and they still have to work a lot in this direction. Therefore, including an Open Data platform as a part of iCity makes a lot of sense.

The cities involved in the iCity project are very advanced regarding their ICT (Information and Communication Technologies) infrastructure and already have -or will have very soon- their own Open Data platform. This document introduces the Fraunhofer Open Data platform, which will be integrated in iCity in order to provide Open Data management functionalities, analysing the already existent platforms in each city and offering an approach for their integration into the iCity platform.

2. Fraunhofer Open Data Platform

The Fraunhofer Open Data Platform is an Open Source Open Data Platform software developed by the Fraunhofer Institute for Open Communication Systems (Fraunhofer FOKUS), located in Berlin, Germany. The first version of the platform has been developed by Fraunhofer FOKUS during the course of the EU Open Cities project (http://opencities.net, Grant agreement: 270896) and it was published under the AGPL version 3 license at GitHub[1]1. Fraunhofer FOKUS is continuing the task of improving the platform and developing new functionalities. The following sub-sections introduce the Fraunhofer OD Platform in the shape of an official detailed documentation.

2.1 Key functionalities

The platform offers an integrated solution for publishing open data. It provides a data portal (i.e., the user front-end), a data registry, and a triple store for RDF-based linked data. The platform is easy to deploy and provides multi-language support. The primary benefit that the platform will offer is a “one-stop-shop” experience, which will enable the development of novel third-party applications. Application developers will have a consolidated view on all of the open data available that has been catalogued, and will allow to navigate, identify, access and use data of interest. The OD platform will serve as the foundation for the pan-European Open Data contest, which will promote the development of novel applications that rely on open data.

2.2 Platform architecture

The OD platform plays a crucial role for its stakeholders, such as the city government councils and open data users. Figure 1 below features a high level architecture diagram of the Fraunhofer OD Platform and its relation to the city infrastructure.

City government stakeholders locally host and maintain city government department websites and content management systems (CMS). These websites and CMS contain data sets that are in the following file formats: csv, html, pdf, xls, xml, etc. In order to enable the accessibility of the data sets over the central Open Data portal, metadata for selected data sets are published using the non-public civil servant portal. The non-public civil servant portal is part of the Open Data portal and requires user authentication.

Relying on web browsers, data users are able to navigate, search, filter and explore the data sets provided by the Open Data portal. Third party (mobile) application developers identify the specific location, where data sets reside, from the corresponding metadata as described in the Open Data portal. Consequently, the applications developed can either directly access the data sets in their original format on the department web sites, or use the SPARQL endpoint in the Open Data portal to access the linked data representation when available (i.e., for select data sets).

The FraunhoferOD Platform was designed using existing software and tools. These include the content management system, the open data registry, the linked data store, and the data enrichment tool.

[1] https://github.com/fraunhoferfokus/opendata-platform

The following sub-sections give a short overview for each of these components.

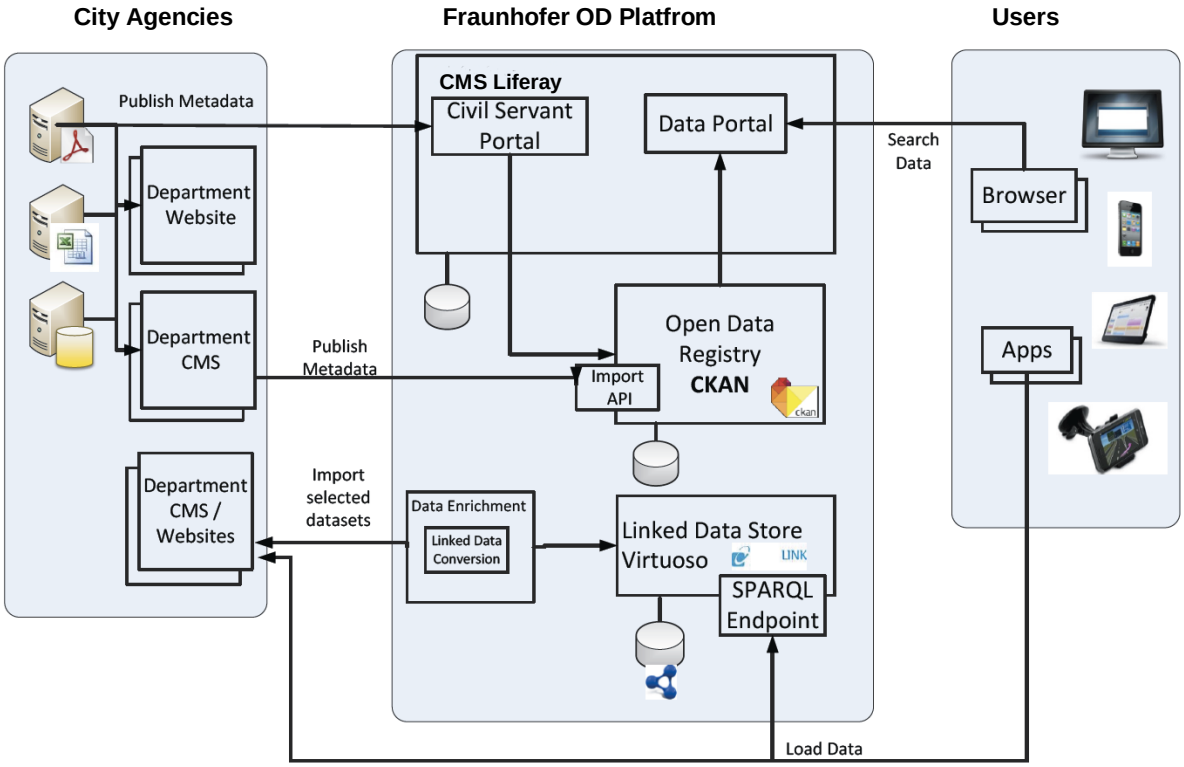

Figure 1: Fraunhofer Open Data Platform High Level Architecture

2.2.1 Data Portal

The Data Portal consists of a publicly accessible part and a non-publicly accessible part. The publicly accessible part providesusers with a web front-end that facilitates data browsing, search, and consumption. Moreover, it provides data access and enables users to discuss matters of interest concerning data sets. The non-publicly accessible part is reserved for civil servants or authorized third parties and allows them to publish metadata to new data sets.

The implementation of the Open Data portal is based on Liferay[2], a portal server by Liferay, Inc. A portal server aggregates several web applications, called portlets, into one web page. Each portlet is developed to handle a specific job. The aggregation of specialized portlets results in a high-scalable modular system.

The Liferay Portal product is available in two versions, namely, the community edition and the commercial enterprise edition. The Fraunhofer OD Platform uses the community edition. Both are licensed under the GNU Lesser General Public License (LGPL).

Liferay features a content management system, a complex permission system and several collaboration tools including a wiki, forum, blogs, instant messaging and email. Since it is open source, every part of the portal is modifiable to one's own specific needs. It's written in Java and it can run on the most popular servlet containers and application servers, such as

[2]http://www.liferay.com/

Apache Tomcat or Oracle Glassfish. In our case, we chose Apache Tomcat for the OD portal.

Figure 2 below depicts a screenshot of the Fraunhofer OD portal start page. As shown here, the user can browse over different data categories or search for datasets.

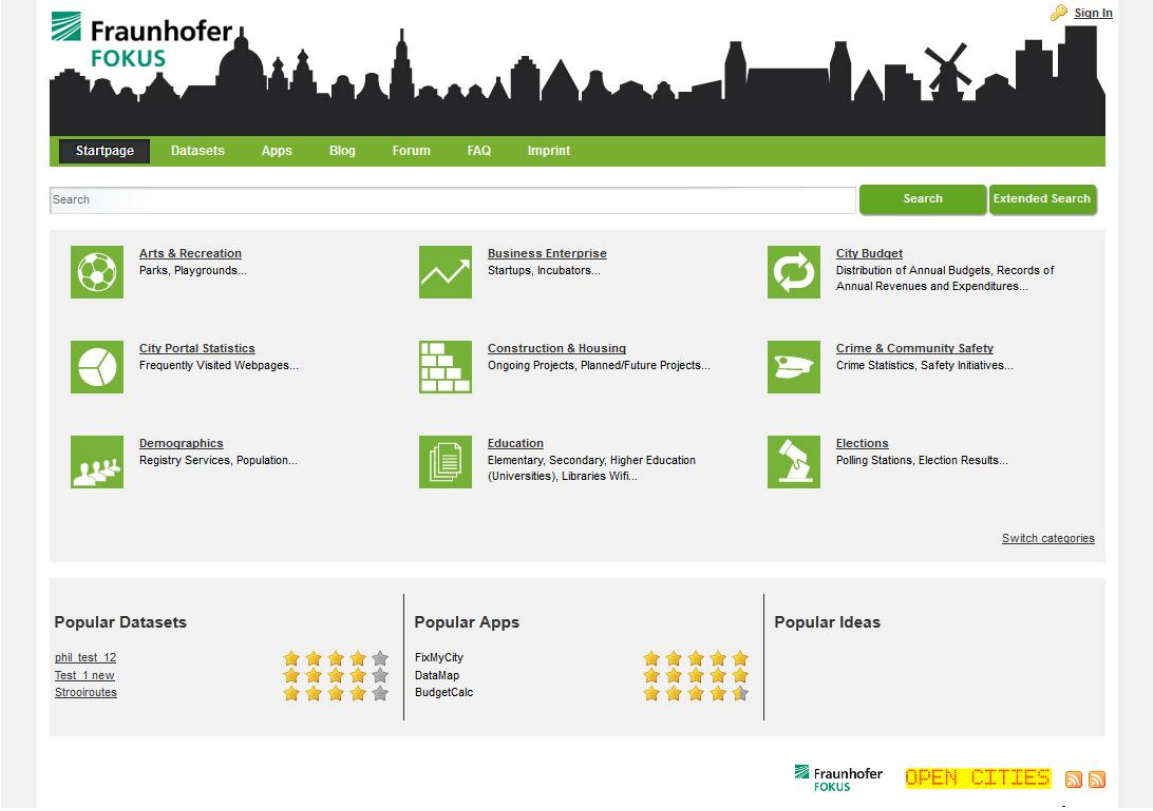

Figure 2: Homepage of the Fraunhofer Open Data Platform

The results of search are presented as a list of short summaries of found datasets as it is depicted in the Figure 3.

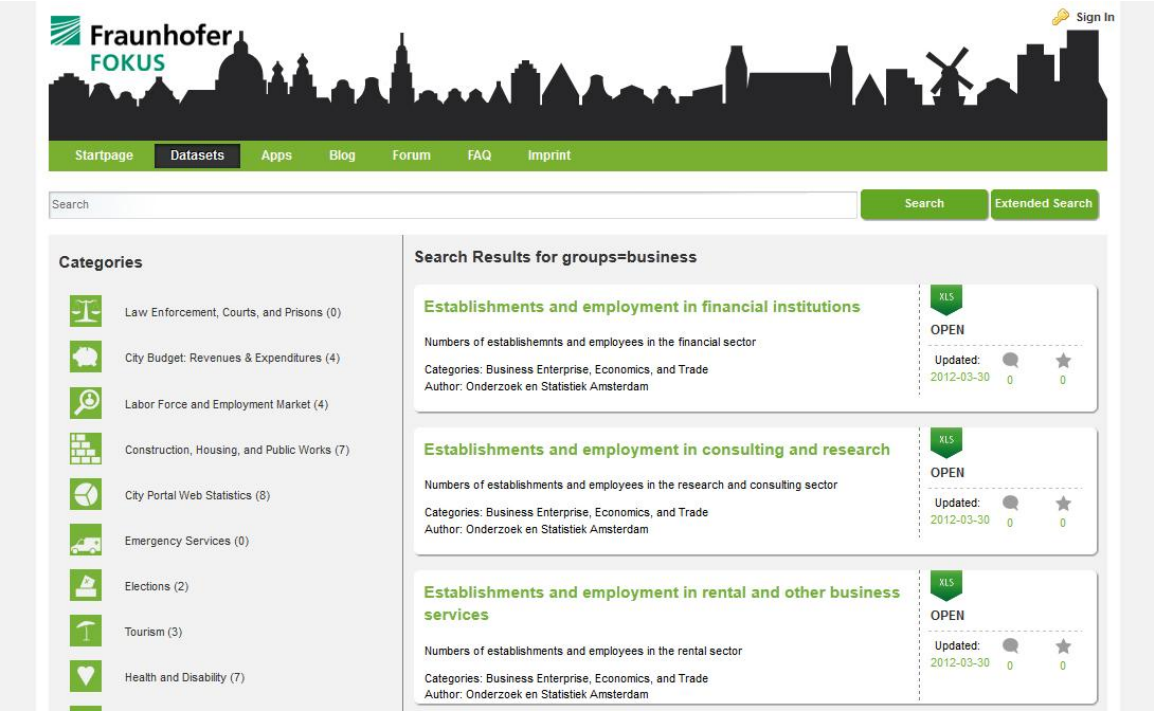

Figure 3:Dataset search results

Users can select one of the found datasets and open a detailed view as it is show in Figure 4 below.

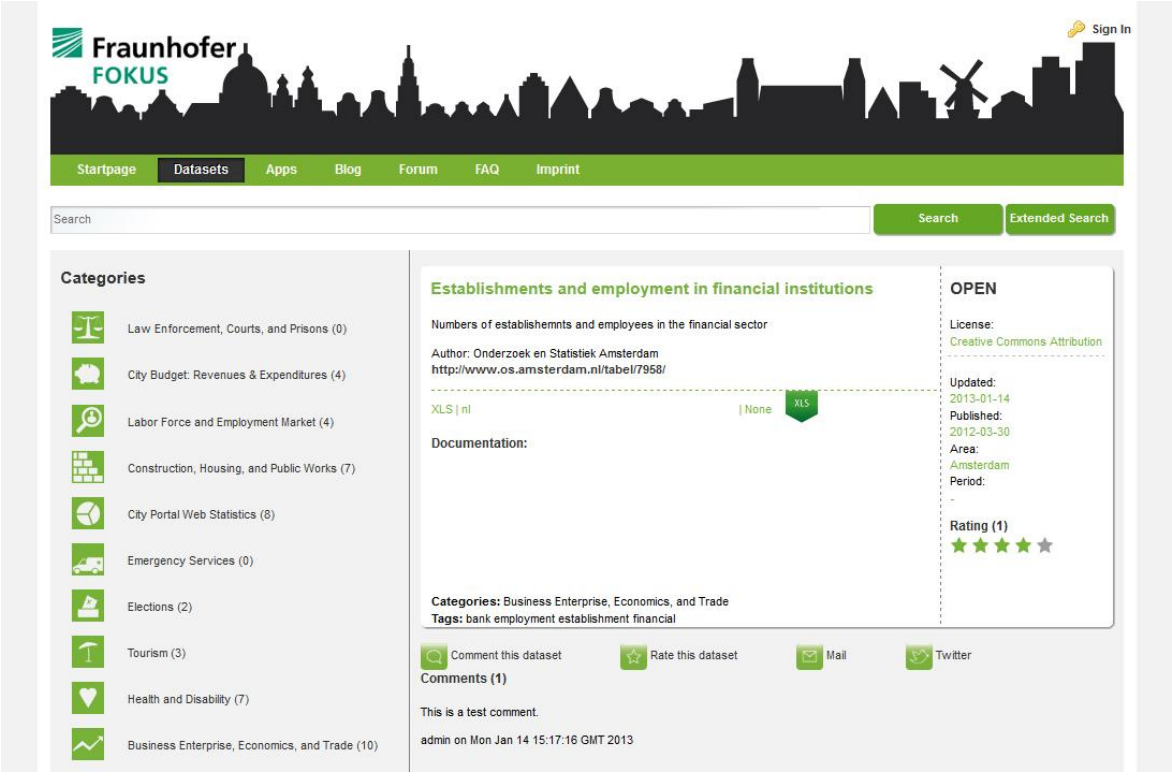

Figure 4: Viewing a Dataset

Representatives of the cities registered in the OD Platform can publish new datasets by filling the dedicated form presented in Figure 5 below.

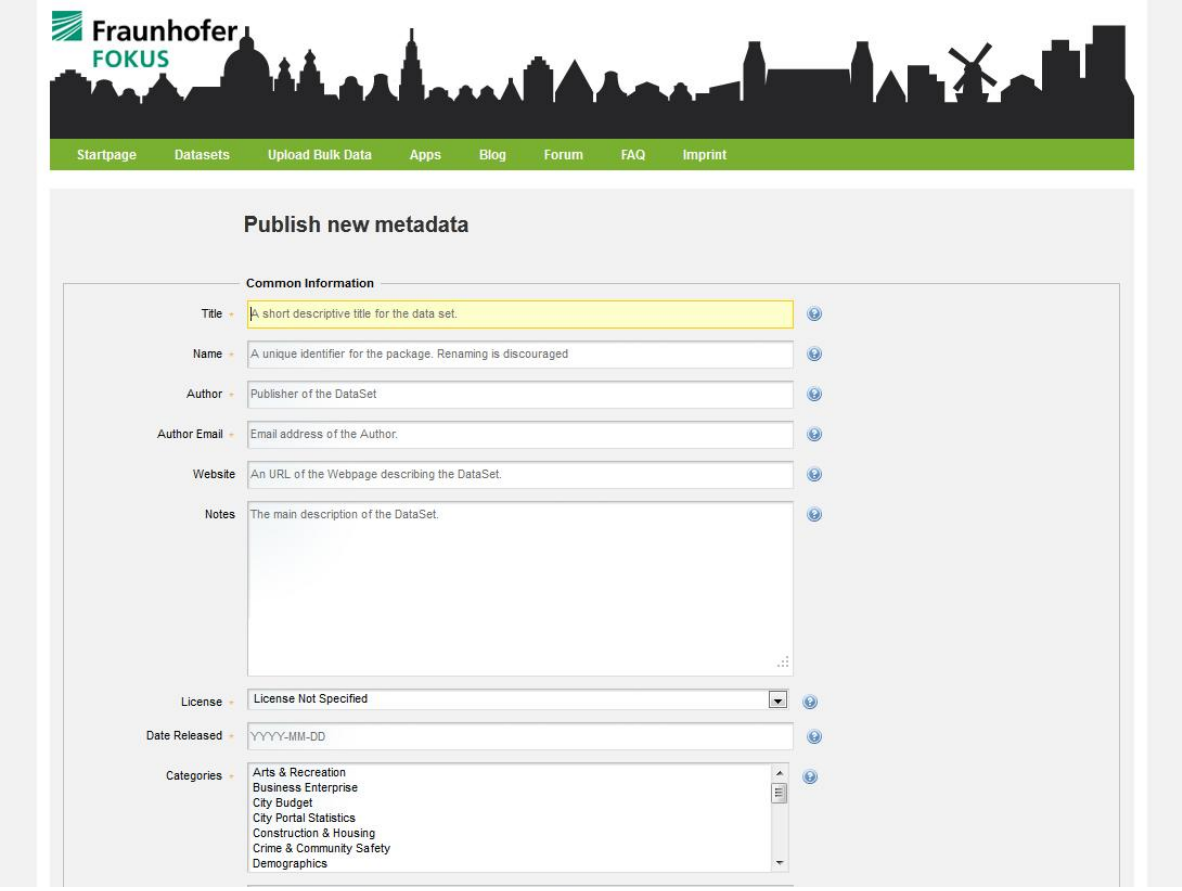

Figure 5: Adding a Dataset

2.2.2 Open Data Registry

The Open Data registry is a major component of the Open Cities OD Platform. It stores the metadata associated with data sets that are catalogued in the data portal. In order to register new metadata, one can either do it by using the Open Data portal or by using the Open Data registry import API, which is accessible via the department CMS and hence publishes the metadata automatically.

Furthermore, the open data registry provides a comprehensive API that allows to read and search through the CMS generated metadata. The Open Data registry is instantiated by using the CKAN[3] (Comprehensive Knowledge Archive Network) data catalogue portal software, the de-facto European standard for metadata registries in the Public Sector Information (PSI) domain.

CKAN is a free software suite maintained by the Open Knowledge Foundation. CKAN is a metadata registry that enables publishing, sharing and finding metadata entries, called data packages. It uses a predefined cataloguing schema built on a set of metadata terms. Among core features of CKAN are customizable metadata registry schema, which enables adding extra fields, and multiple ways for maintaining metadata entries (i.e., via the CKAN admin web page or via the CKAN API). Open interfaces of CKAN enable seamless integration and federation with other open data portals.

[3]www.ckan.org

2.2.3 Linked Data Store

Besides providing references to metadata for data sets hosted on department's websites, the Open Data portal also provides a service to host linked data for select data sets. Linked data enables the seamless aggregation of data sets from different and heterogeneous origins and offers powerful filtering and querying mechanisms. Linked data are physically stored adhering to the RDF in the linked data store of the Open Data portal. Machine readable access to be used by third party (mobile) applications and mashups is provided via the query language for RDF (SPARQL) endpoints.

The Linked Data Store is implemented on top of OpenLink Virtuoso[4]. Virtuoso is used to deploy and access linked data, which is provided in RDF files. A third party or an application can access the data over a SPARQL endpoint. Moreover, SPARQL provides the possibility to make a simple query over different datasets/graphs, so different datasets can easily be aggregated with one request.

Virtuoso offers the possibility to load large datasets into one or more graphs. It also integrates user and rights management like most DBMS (database management systems). Virtuoso has a web interface, where most of the configuration can be done. Once a data set is installed into a graph, the SPARQL endpoint can be queried over a web interface. Another possibility is to query the endpoint via an HTTP request, after which the system can deliver the results in different formats like HTML, JSON, XML or even JavaScript.

2.2.4 Data Enrichment

Data enrichment describes a conceptual component used for linked data conversion, i.e., transforming CSV files into RDF files described with a defined ontology. Semantically sound data enrichment requires manual efforts and can be supported using tools. The DERI[5] RDF Extension for Google Refine is used to assist in the provisioning process of selected data sets and their conversion into linked data.

2.3 Metadata structure

The metadata schema for the FraunhoferOpen Data Platform was defined by extending the CKAN metadata schema. The detailed specification of metadata structure is presented in the Annex I, section 9.2.

[4] http://virtuoso.openlinksw.com/

[5] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

3. OD Platform in Barcelona

OpenData BCN is the open data platform of the city council of Barcelona. It provides public sector information for everybody in standardized and digital formats, so it can be understood by everybody. The main aim of the project is to encourage people to reuse the information and create new services with it. It does not matter if it is used for social or commercial projects.

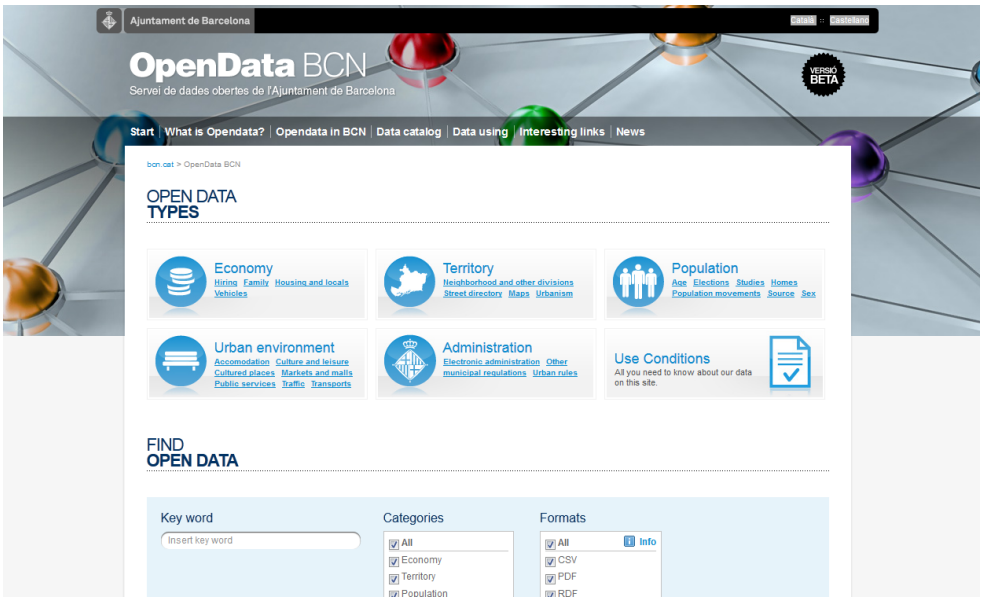

Figure 6: Homepage of OpenData BCN

At present the platform offers an easy web access to the data catalogue. Each dataset is sorted in categories and tagged. The portal supports the following categories/subcategories:

- Economy
  - Hiring
  - Family
  - Housing and locals
  - Vehicles
- Population
  - Age
  - Elections
  - Studies
  - Homes
  - Population movements
  - Source
  - Sex

- Administration
  - Electronic administration
  - Other municipal regulations
  - Urban rules
- Territory
  - Neighbourhood and other divisions
  - Street directory
  - Maps
  - Urbanism
- Urban environment
  - Accommodation
  - Culture and leisure
  - Cultural places
  - Markets and malls
  - Public

The data itself is appended in one or more different formats. Users can simply search for a key word or/and apply several filters for categories (but not for subcategories) and formats (CSV, PDF, RDF, XLS, XML, Other). Furthermore, it is possible to post a message about a selected dataset via different social platforms like Facebook and Twitter. The key characteristics of the platform are presented in the table below.

URL	**w20.bcn.cat/opendata**
Number of datasets	428 (10.01.2013)
Technology	ASP.NET
Resource Formats	CSV, PDF, RDF, XLS, XML, Other
State	Online in Beta Version
Languages	GUI - English, Spanish, Catalan Metadata – only in Catalan
Meta Data Structure	Data, Categories, Description, Tags, Publication date, Last update, License, Data Source

API:

The platform does not provide a sophisticated API for developers. Instead, it is possible to download the entire catalogue as an RDF file[6]. By parsing the received XML file an easy access to the entire data catalogue can be achieved. Every data field and links to the

[6] http://w20.bcn.cat/opendata/CatalegRDF.aspx

resources are included. The metadata information stored in the OD Platform and accessible through the catalogue is only in Catalan.

4. OD Platform in Bologna

The Open Data Platform of Bologna was launched to enable companies, associations and citizens to use and evaluate the public data, and develop new applications. The datasets can be filtered by category, tag and year, or searched by a key word. The search functionality works on both the data catalogue and the entire content of the platform. Platform users can rate and comment datasets without registration. The platform is still under development and only some elements of the portal are translated into English. Practically the entire content of the platform including data and metadata is on Italian. A screenshot of the main page and a short summary of its key characteristics can be foundbelow.

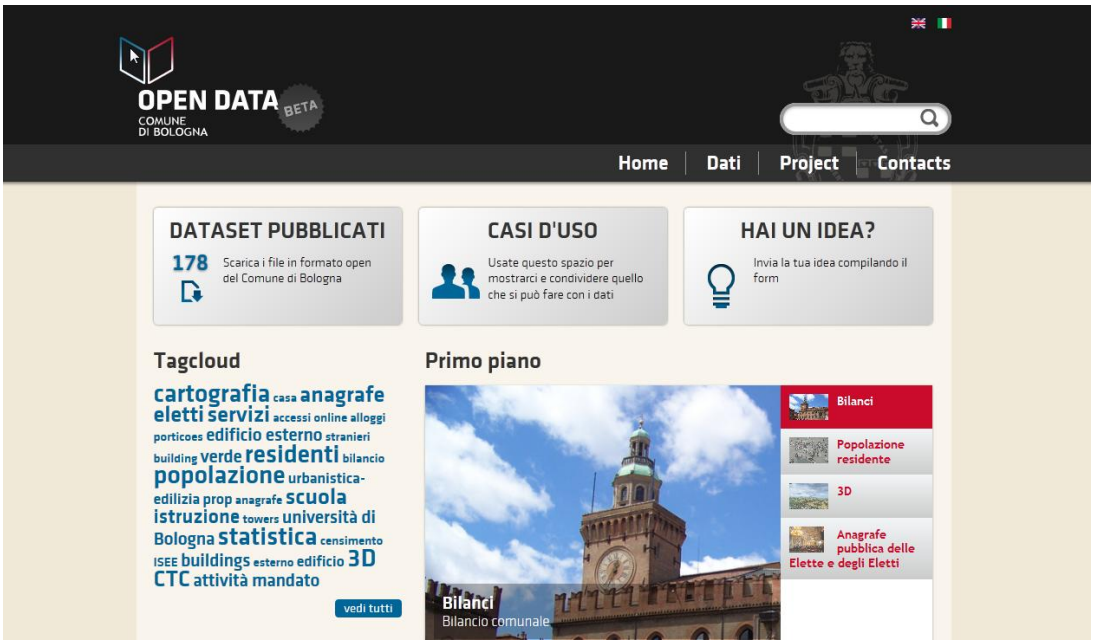

Figure 7 Homepage of Open Data Comune Di Bologna

URL	**http://dati.comune.bologna.it/**
Number of datasets	178 (10.01.2013)
Technology	Drupal (for the portal), MySQL database (for the backend)
Resource Formats	Any type of files. Mostly contains CSV, ZIP files
State	Online in Beta Version
Languages	Italian, English(only some elements of GUI)
Meta Data Structure	Description, Resources, Issue, Rating, Responsibility, Tags, Source, License, Comments

API:

The platform offers a RSS feed[7] containing all metadata stored in the platform. The feed is updated on adaily basis. For each dataset, the feed contains the following information:

[7] dati.comune.bologna.it/tuttidati.xml

*<title>*Dataset title*</title>*

*<description>*Dataset description*</description>*

*<category>*Dataset category*</category>*

*<enclosure>*Link to the corresponding data file*</enclosure>*

*<guidisPermaLink="true">*Address of the dataset in the OD Portal, for example:http://dati.comune.bologna.it/node/320*</guid>*

*<pubDate>*Date of dataset publication*</pubDate>*

*<opendatacobo:version>*Version of the dataset*</opendatacobo:version>*

*<opendatacobo:license>*Dataset license*</opendatacobo:license>*

*<opendatacobo:coverage>*Time coverage (year)*</opendatacobo:coverage>*

5. OD Platform in Genoa

The city of Genoa is currently developing an open data platform. It is planned that the platform will be available online in February 2013. The platform is developed on top of the CKAN[8] platform, the currently leading European technological platform for Open Data repositories. The main drawback of CKAN is its relatively limited representation layer. Therefore, in the Genoa Open Data Platform,CKAN is used only as a metadata storage.The portal will be implemented on the basis of Drupal CMS[9].

URL	Will be available at: **http://dati.comune.genova.it**
Number of datasets	0
Technology	CKAN, Drupal and MySQL
Resource Formats	Not specified yet
State	Under Development
Languages	Italian
Meta Data Structure	Standard CKAN Data Structure

API:

The open-source data portal platform CKAN offers an extensive API to access the data catalogue. The entire API is based on JSON and provides a RESTfulWebservice, which maps every possible operation on the data catalogue. These include reading, creating, deleting and updating datasets. To perform a search on the data catalogue a simple REST call of `[URL]/api/3/action/package_search?q=keyword`leads to an easy readable JSON result. It is also possible to perform complex search queries by using SQL via the API. Detailed information about the entire API can be found in the official documentation: http://docs.ckan.org/en/latest/datastore-api.html

[8]www.ckan.org

[9] http://drupal.org/

6. OD Platform in London

The London Datastore is a project held by the Greater London Authority (GLA). It provides data from the GLA and other London public sector organizations. Citizens are encouraged to use that data in any way. Furthermore,reusing the data in meaningful ways in apps, websites or mobile products is encouraged. There are several ways to browse the data catalogue. The datasets are sorted into categories, keywords and organizations. A searchfunctionality is offered as well. Each dataset comes with several metadata, a rating feature and the possibility to post comments.A screenshot of the datastore and short descriptions of its characteristics can be found below.

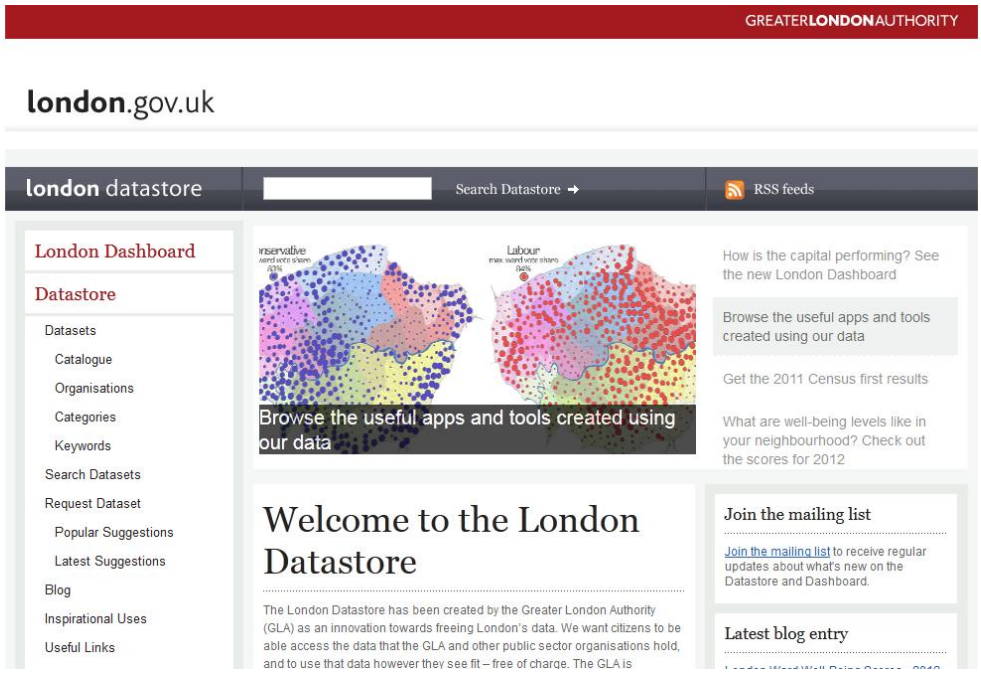

Figure 8 Homepage of the London Datastore

URL	data.london.gov.uk
Number of datasets	563 (14.01.2013)
Technology	Drupal
Resource Formats	HTML
State	Online
Languages	English
Meta Data Structure	See Chapter 9.1

API:

The entire data catalogue is accessible as a CSV file[10]. This file is updated daily from the MySQL database, which holds the actual data. It includes full metadata of every dataset, the description of which can be found in the Annex I of this document.

[10] http://data.london.gov.uk/datafiles/datastore-catalogue.csv

7. iCityOpen Data Platform Integration

The functionalities of the Open Data management will be provided by the iCity platform by integrating it into the Fraunhofer OD Platform, which will be called from now oniCity OD Platform in the context of iCity. As a component of the overall system, it will stay relative loose coupled, sharing with other components only the iCity user management service. In the Fraunhofer OD Platform, the user management is implemented on top of the user management functionality in the Liferay CMS. For integration with the iCity, the user management in the Liferay CMS instance should be modified, so that it uses the dedicated iCity user management service instead of an internal one. This will enable centralised user management in the iCity Platform.

As it was presented in the previous chapters, the cities involved in the project already have their own OD platform. Their common interest is the integration of their OD platforms on top of iCityin order to be able to benefit of the cross-platform search.That will enable iCity users to easily find open datasets in the cities through a common search interface and will foster the use of the published Open Data. The iCity OD (Fraunhofer OD) platform with its rich metadata and rich set of search functionalities,is well suited to be a basis for the implementation of this service. Figure 9sketches the implementation of the cross-platform open data search functionality.

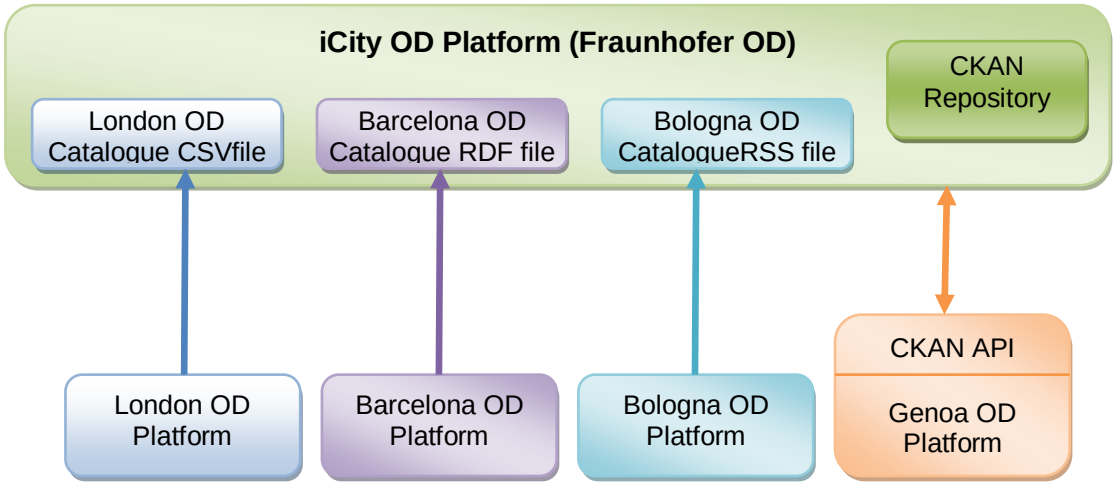

Figure 9:OD cross-platform search implementation concept

The idea for the implementation of a cross-platform search is very simple. The iCity OD Platform will download the OD catalogues of London OD, Barcelona OD and Bologna OD in a daily basis, and will parse and convert them into the iCity OD metadata structure, placing them later in a temporary CKAN repository. The iCity OD Platform will provide an advanced metadata based search functionality, which will search for datasets integrated in its CKAN repository, which is the foundation of the iCity OD. Then, it will search through the temporary CKAN instance containing metadata from the cities and, finally, it will perform a direct search request to the CKAN API of the Genoa OD Platform. The collected data will be presented in the iCity OD Platform.

8. Conclusion

The Fraunhofer OD Platform will be integrated into the iCity Platform as a service to provide Open Data management functionalities. As it was presented in the previous sections of this document, London, Barcelona, Genoa and Bologna already have their own Open Data platform or they are at the latest stage of its development (in Genoa). From all mentioned platforms, the Fraunhofer OD has the richest metadata structure and the most advanced functionalities. However, these cities have already invested a huge amount of resources into their development and deployment, and asides from London, they do not plan to migrate to Fraunhofer OD or any other platform. Also, none of the cities is interested in duplicating their Open Data. London is interested in migrating from their outdated platform into a newer, more advanced one, and the Fraunhofer OD Platform, as part of iCity Platform, is an interesting option for them.

Consortium partners have held many discussions on how the existing OD platforms of each city have to be integrated into iCity. In the end, they concluded that implementing the cross-platform OD search on top of the Fraunhofer OD Platform integrated in the iCity is the most viable scenario, adding value to users. This scenario will be implemented during the next stage of the project, when the Bologna OD Platform will already be available.

The Open Data Platform is an important component of the iCity Platform, since its future exploitation is not limited by the above mentioned cities, and it provides functionalities required in all European cities, some of which may be interested in deploying it as part of the iCity service package.

9. Annex I: Metadata Structures

9.1 London Datastore Metadata structure

DRUPAL_NODE	Internal Drupal Node
TITLE	Title of the Dataset
DATASTORE_URL	URL of the Dataset package on the London Datastore site [metadata]
DDATE	Creation Date of the Dataset (in the London Datastore)
CATEGORIES	One or more dataset categories for the package
LONGDESC	Description of the Dataset
GEOGRAPHY	Geographical level e.g. ward
EXTENT	Geographical extents e.g. Greater London
AUTHOR_NAME	Name of the Dataset Author
MAINTAINER	Name of the Dataset Maintainer (or sub-name of Dataset Author)
MAINTAINER_EMAIL	Email address of the Dataset Maintainer
UPDATE_FREQUENCY	How often the Dataset is updated
DATE_RANGE	Date range the Dataset covers
RELEASE_DATE	Release Date of the Dataset (by the Dataset Author)
METADATA_UPDATE	Last Updated Date of the Dataset or metadata (in the London Datastore)
LICENSE_SUMMARY	Summary of the license applicable to the Dataset
URL	URL of link to further information
DOWNLOAD_URL	Download URL to the Dataset
EXCEL_URL	Download URL to the Dataset in MS Excel format
CSV_URL	Download URL to the Dataset in CSV format
XML_URL	Download URL to the Dataset in XML format
GOOGLEDOCS_URL	Download URL to the Dataset on Googledocs
JSON_URL	[Possible future format]
SHP_URL	Download URL to the Dataset in ESRI Shape File
KML_URL	[Possible future format]
TAB_URL	Download URL to the Dataset in MapInfo Tab File
APPS_USING_THIS_DATA	Names of developed Applications or Inspirational Uses making use of this Dataset
TAGS	Keywords
apps_using_this_data_urls	URLs of developed Applications or Inspirational Uses making use of this Dataset
license_details	Details of the License applicable to the Dataset
spatial_ref	Spatial reference of data if geographical e.g. British National Grid

9.2 Fraunhofer Open Data Platform Metadata structure

In Table 1, we list the metadata fields for the metadata schema, including a short description, format of the field, and whether field is mandatory or not.

Metadata Field ID (* denotes mandatory)	Explanation	Format

title*	explanatory, human friendly title	string
name*	machine friendly identifier, not to be changed	string (only 'a-z0-9' and '-_')
Notes	more detailed description of the dataset	string
groups*	categorization into fixed classes	list of Strings
Tags	keywords	list of Strings
author*	publishing institution or department	string
author_email*	contact person email address for a dataset	string: Email
Maintainer	data portal user (i.e., individual who publishes)	string
maintainer_email	respective email address	string: Email
Version	release version of the dataset	string
license_id*	license	string (e.g., cc-by or closed)
URL	website with more details	string: URL
State	“active”: dataset is complete and published, “deleted”: dataset is deactivated and not visible any more	string
extras:date_released*	release date	string: YYYY-MM-DD
extras:date_updated	date of last update	string: YYYY-MM-DD
resources*	list of resources	list of JSON-Dictionaries
resources :url*	direct link to file or service	string: URL
resources :format	format of resource	string: file extension or MIME-type
resources :description*	description	string
resources:language*	language of the resource	two letter ISO language code
resources:url_doc	link to doc for resource	string: URL

resources:steward	if this resource was uploaded not by the publisher, but by a data steward, i.e. a third party that converted a CSV into an RDF file, that person can be referenced here, e.g. with an URI	string
extras:geographical_coverage	geographical coverage	string
extras:geographical_granularity	geographical granularity	string
extras:temporal_coverage-from	start of period covered	string:YYYY-MM-DD
extras:temporal_coverage-to	end of period covered	string:YYYY-MM-DD
extras:temporal_granularity	temporal resolution in sec	integer
extras:ratings	quantitative feedback of users on packages	list of dictionaries
extras:rating:date	date rating was issued	“<day of the week><month><day of the month><hours>:<minutes>:<seconds><timezone><year>” e.g. “Fri Oct 07 14:29:38 CEST 2011”
extras:rating:userId	rating author	string
extras:rating:ratingValue	rating value	int
extras:comments	qualitative feedback of users on package	list of dictionaries
extras:comment:date	date comment was issued	“<day of the week><month><day of the month><hours>:<minutes>:<seconds><timezone><year>” e.g. “Fri Oct 07 14:29:38 CEST 2011”

extras:comment:userId	maintainer	string
extras:comment:comment	comment	string

Table 1: Open Data Platform Metadata Schema

The CKAN metadata field <group> is mapped to data categories in the OD platform. The complete list of data categories in presented in Table 2

Category Name	Group Identifier
Arts and Recreation	rec
Business Enterprise, Economics, and Trade	business
City Budget: Revenues & Expenditures	budget
City Portal Web Statistics	stats
Construction, Housing, and Public Works	housing
Crime and Community Safety	safety
Demographics	demographics
Education	edu
Elections	elections
Emergency Services	emergency
Energy and Utilities	energy
Environment, Geography and Meteorological	environment
Health and Disability	health
Labor Force and Employment Market	employment
Law Enforcement, Courts, and Prisons	law
Political	politics
Tourism	tourism
Urban Transport	transport
Others	misc.

Table 2: List of Data Categories

In Table 3, is presented the list of data licenses that arise in the OD platform.

License ID	License Title
apache	OSI Approved::Apache Software License

apache2.0	OSI Approved::Apache License, 2.0
bsd-license	OSI Approved::New and Simplified BSD licenses
ca-tosl1.1	OSI Approved::Computer Associates Trusted Open Source License 1.1
cc-by	OKD Compliant::Creative Commons Attribution
cc-by-sa	OKD Compliant::Creative Commons Share-Alike
cc-nc	Non-OKD Compliant::Creative Commons Non-Commercial (Any)
cc-zero	OKD Compliant::Creative Commons CCZero
eclipse-1.0	OSI Approved::Eclipse Public License
gfdl	OKD Compliant::GNU Free Documentation License
gpl-2.0	OSI Approved::GNU General Public License (GPL)
gpl-3.0	OSI Approved::GNU General Public License version 3.0 (GPLv3)
mit-license	OSI Approved::MIT license
mozilla	OSI Approved::Mozilla Public License 1.0 (MPL)
mozilla1.1	OSI Approved::Mozilla Public License 1.1 (MPL)
notspecified	Other::License Not Specified
odc-odbl	OKD Compliant::Open Data Commons Open Database License (ODbL)
odc-pddl	OKD Compliant::Open Data Commons Public Domain Dedication and Licence (PDDL)
W3C	OSI Approved::W3C License

Table 3: List of Data Licenses