European Public Sector Information Platform

Topic Report No. 2016/03

UNDERSTANDING

THE EUROPEAN DATA PORTAL

Published: February 2016

Table of Contents

Keywords

EDP; European Data Portal; pan-European; portal; open data; reuse

Abstract/ Executive Summary

This report presents the technical arquitecture of the European Data Portal. This portal, available since 16 November 2015, aims at being the new one-stop shop on PSI re-use in Europe, taking over all the promotion duties of ePSI Platform. The European Data Portal includes information, tools, and a multilingual data catalog where all the open datasets from the EU Member States are being federated to create a single entry point for European-wide stakeholders who want to find and reuse PSI in Europe.

Introduction

More and more volumes of data are published every day. The amount of data across the world is increasing exponentially. A substantial amount of this data is collected by the public sector. But for the data to be re-used, it needs to be accessible.

The Beta version of the European Data Portal is available since 16 November 2015. The **Portal harvests the metadata available on public data and geospatial portals across European countries**. Portals can be national, regional, local or domain specific. They cover the 28 EU Member States, EFTA countries and countries involved in the EU's neighbourhood policy.

But how does the Portal function? This factsheet provides a summary of the architecture of the European Data Portal.

The Architecture

For a better understanding of the integration of the components into the overall architecture, each components' functionality as well as interactions from different perspectives (user/system) are described below.

The following figure provides a high-level overview diagram of the European Data Portal architecture.
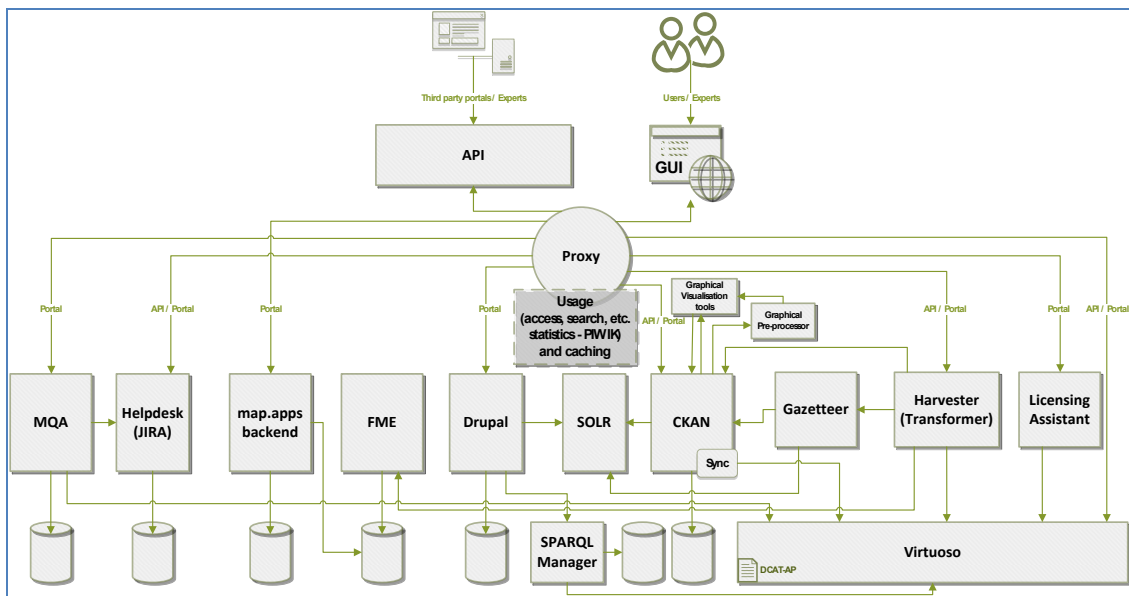
Access to the Portal

The access to the Portal is provided in two ways: **a machine-readable API** and **a human readable web site** –Graphic User Interface or GUI. The API enables its users to search, create, modify and delete metadata on the portal.

The GUI is basically built on two components: **CKAN** and **DRUPAL**. CKAN manages and provides metadata content (datasets) in a central repository. DRUPAL provides the Portal's Home Page with editorial content (e.g. Portal's objectives, articles, news, events, tweets, etc.) and links to an **Adapt Framework**-based training platform. In addition it offers extended functionalities to registered users via user login.

Both systems are used in a side-by-side architecture. A **proxy** is responsible for delivering the

web pages requested by the user. Both systems are equally themed with the same Look&Feel so that the user is not aware on which system he/she is currently browsing.

Searching the Portal

The portal uses the **SOLR search engine** in order to separately search for editorial content in DRUPAL and for datasets in the CKAN repository. The GUI includes a **Licensing Assistant** component that supports the user by providing legal information on the usage of a specific dataset in terms of licenses that apply to the dataset.

The **SPARQL Manager** component allows the user to enter and run SPARQL queries on the **Virtuoso** linked data repository. It also allows the logged-in user to store and re-run SPARQL queries and notifies the user when a query has finished running.

Geo spatial data

Using the **map.apps** backend application, **Geo spatial data** is visualised on geographical maps. The application is a proprietary solution that comes with different tooling and thematic focus, a graphical configuration interface, supports responsive web-design and internationalisation files.

The application also implements the OSGI specification on the client side (in JavaScript) allowing sharing and re-usage of the bundled application logic as well as a straightforward maintenance.

Statistical data that is linked to datasets can be visualised in tabular (tables) and graphical (charts) form using a D3.js library.

Harvesting Data

On the Harvesting side, the portal follows a two-fold architecture too. **CKAN** is used as the central metadata repository for storing, browsing and searching datasets in a **POSTGRES** relational database.

In order to also support a linked data functionality the CKAN metadata is replicated into a Virtuoso quad store repository via a CKAN synchronisation extension, in order to ensure that both repositories have the same set of metadata.

The Harvester is a separate component that is able to harvest data from multiple data sources with different formats and APIs. The harvester is acting as a single point of entry for all metadata that gets harvested, transformed into the CKAN JSON schema and pushed into the CKAN repository.

The Gazetteer component is used by the Harvester to enhance the metadata with geo-spatial data and information (geo-coordinates, names, places, etc.). The Gazetteer is mainly used to improve the search functionality. It uses the **FME** component as a universal spatial ETL (Extract-Transform-Load) tool that supports the accessing, processing and outputting of all spatial file/database formats and that is used for harvesting the sources for geographical names.

Enhancing quality

The Portal architecture includes three additional components to enhance the quality of the metadata and the portal. A **Helpdesk** handles user support requests and feedback.

The **Metadata Quality Assistant (MQA)** periodically generates reports on the quality of the harvested metadata and creates tickets for the Helpdesk in case of harvesting issues.

The third component is the **monitoring component based on PIWIK** and located at the Proxy in the architecture. In the full respect of data privacy, it records requests and user interactions on the portal in order to generate anonymised user traffic statistics that will help enhancing the usage of the Portal.

Summary of Components

COMPONENT	DESCRIPTION
API Access	Machine-readable (SOAP / REST) API
GUI Access	Portal website graphical user interface
CKAN	Portal's central metadata (dataset) repository
DRUPAL	Portal's Home Page managing editorial content
ECAS	European Commission Authentication System used for user registration and login in order to provide extended functionalities of the Portal
Adapt Framework	Online platform used for Portal Training Modules (available in EN + FR)
Proxy	Routing of (HTTP(S)) user requests to corresponding components
SOLR Portal Search (editorial data)	Search engine used for searching portal editorial content
SOLR Dataset Search (metadata)	Search engine used for searching and filtering datasets in the CKAN metadata repository
Licensing Assistant	Component to provide legal information on (re-)usage of specific datasets
SPARQL Manager	SPARQL query editor allowing to run SPARQL queries on linked data in the Virtuoso repository
Virtuoso	Linked data quadruple store that is synchronized with the CKAN repository
map.apps: Geo-spatial Data Visualisation	Proprietary application to visualize geo-spatial data and information using geo-maps. It comes with different tooling and thematic focus, a graphical configuration interface, supports responsive web-design, i18n internationalization files, client side implementation of the OSGI specification (JavaScript)
Graphical Data Visualisation	Recline.js/D3.js JavaScript libraries to visualize (statistical) data in tables and graphical charts
Pre-processor	RESTFUL web API, running on a Node.js server which analyzes and transforms XSL/XLSX files into CSV format in order to be used from the Visualisation tool
Harvester	Single entry point component for harvesting data from multiple data sources in different formats and from different APIs
Gazetteer	Component providing geo-spatial data and information
FME	Component used by the Gazetteer as a universal spatial ETL tool (Extract-Transform-Load) that supports the accessing, processing and outputting of all spatial file / database formats and that is used for harvesting the sources for geographical names
smart.finder	Component used by the Gazetteer and simplifying searches for spatial data, services and documents. It enables fast and structured access to extensive, distributed and heterogeneous data stores
Helpdesk	Portal offers a user request/feedback form via the JIRA-API and generates JIRA tickets for follow-up by helpdesk
Metadata Quality Assistant (MQA)	Component to report on the quality of the harvested metadata and to alert helpdesk in case of issues

Monitoring	PIWIK component that provides Traffic Analytics of portal usage
Multilingual Support	Web pages + core editorial content + dataset descriptions available in all 24 official EU languages
MT@EC	Machine Translation Services of the European Commission used for translation of the metadata into all of the supported languages by the portal

For more information, please visit the [European Data Portal](#) or contact via [email](#).

The European Data Portal initial content has been collected by harvesting national public data and geospatial portals. Progressively, the portal will harvest additional metadata collected from regional, local and domain specific portals. **Do you want your portal or website to be harvested by the European Data Portal?** Read the requirements.

Share your story about how you make use of Open Data. Are you an entrepreneur? A non-governmental organisation? A civil servant responsible for publishing data? A local authority? Tell us your story! The purpose of the collection of use cases is to assemble interesting European stories about the benefits and efficiency gains that result from the use of Open Data.

About the Consortium

Capgemini Consulting
www.capgemini-consulting.com

ODI open data institute
www.opendatainstitute.org

INTRASOFT INTERNATIONAL
www.intrasoft-intl.com

time.lex
www.timelex.eu

SOGETI
www.sogeti.com

UNIVERSITY OF Southampton
www.southampton.ac.uk

con·terra
www.conterra.de

Fraunhofer FOKUS
www.fokus.fraunhofer.de

Copyright information