Epimorphics
CONNECTING DATA

Linked Data Publishing Platform (LDPP)

Document: GCloud/ServiceDescription/LDPP/1
Version: 3.0 Epimorphics Ltd
Date: 18th May 2019
Contact: gcloud@epimorphics.com

Service Overview

Summary

The Linked Data Publishing Platform (LDPP) - our main data sharing platform - is a resilient, scalable and cloud-based solution for publishing data. It hosts data for a wide range of our public and private sector clients, and is available as a fully hosted and managed enterprise solution, or elements can be installed locally on your own infrastructure. It is widely used for publishing linked data on data.gov.uk, including data at: landregistry.data.gov.uk, data.food.gov.uk, environment.data.gov.uk, and many others.

Organisations use the the platform for providing data through developer friendly RESTful APIs for internal integration, improved data management, opening data for use. This can be real and near real-time data, static or reference data. We've helped many organisations make their data interoperable supporting open standards and the Open Data principles.

Features

Fully standards-compliant linked data publication platform

Widely used within the UK public sector including for near real-time data

Replicated for fault-tolerance and scalability

Entry level service also available

Benefits

Robust, reliable publication of sustainable, trusted and usable 5-star data

Adaptive and flexible platform - can be grown to meet your changing needs

Deployment flexibility on the cloud or within your own infrastructure

Provide your customers with integrated, live updates twenty-four hours a day

Helps you to build data integration into your transforming organisation

Supports the publication of data that is actually used

Introduction to Epimorphics

Founded in 2009, Epimorphics are at the forefront of developing standards, tools and applications, working with organisations to organise and publish data. We help organisations connect data, making more effective use of your data assets through open, standards-based linked data tools and techniques. Our work includes:

- Over 100 projects for over 30 customers
- 10s of 1000s of people using our systems every day
- 100s of live data collections maintained with updates from hourly to monthly

We have fifteen employees and are actively recruiting. We do not follow a rigid hierarchical structure but put together teams from across the organisation to meet the needs of specific projects. We do not have a separate support team and any employee may be involved in handling support requests. For each customer we can provide a dedicated support email address which links to a suitable group of Epimorphics experts.

Our work

We have around 20 active customers and provide managed services to around 10 different organisations. We have developed linked data solutions for both the public and private sectors and have clients worldwide. We build solutions working with our clients using user centred design techniques. We are proud of our work and the great feedback we receive, for example:

> *"with Epimorphics we get trusted advice helping us to develop and deliver our strategy. They have been working closely with us as a valued partner, helping us to deliver our data driven organisation vision. As a result, our data is more accessible, usable, and enables us to better deliver our priorities."* **-- a Director at one of our public sector clients**

> *"Epimorphics have been working closely with us as an agile supplier, helping us to deliver our open data commitments and vision of changing the way that we provide information on the environment. As a result, our data is more open, richer, and better able to support innovative reuse."* **-- a Service Manager at one of our public sector clients**

> *"Working with Epimorphics is easy. You are adaptable, personable and professional. You deliver a very high quality service that offers good value for money. You don't make it hard to be a customer."* **-- a Service Owner at one of our public sector clients**

Our team has a wealth of experience in all aspects of Linked Data and semantic technologies. We are recognised globally for building and contributing to the development of open data standards and open source technologies that are trusted and used in the public and private sector. Our team have also been active in supporting the development of Open Data Initiatives within the UK. We helped to develop the standards that underpin linked data. We co-developed Apache Jena, the most widely used semantic web development environment, and ELDA, our open source implementation of the Linked Data API. We aim to design solutions that balance technical and business requirements sustainably.

Platform Details

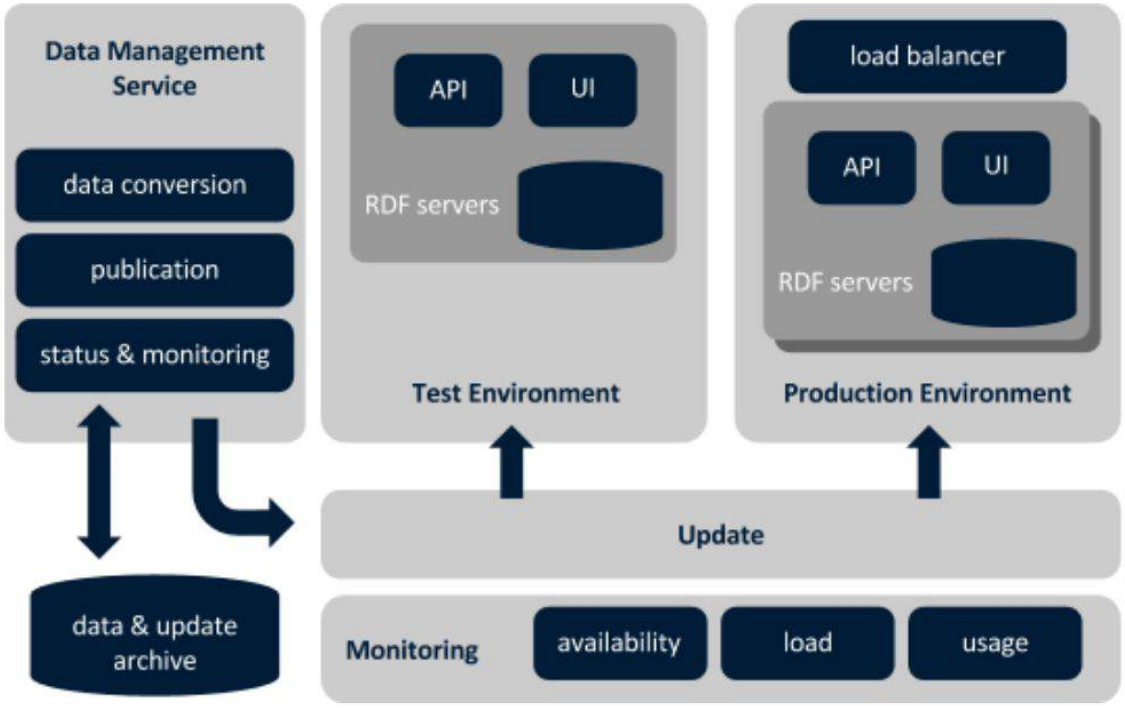
The Linked Data Publishing Platform (LDPP) is widely used for publishing linked data on data.gov.uk, including data at: **environment.data.gov.uk**, **data.food.gov.uk**, **landregistry.data.gov.uk**, and many others.

We offer the platform as a fully hosted and managed service for publishing linked data; by default we provide a hosted service on top of Amazon Web Services.

The platform includes:

- A Linked Data API engine, providing access to the data in several developer-friendly formats (including JSON and CSV) and human-readable web pages
- Customisable text search
- A triple store for storing data as RDF
- A fully SPARQL 1.1-compliant endpoint
- A scale-out, fault tolerant runtime platform
- A data management system, to enable clients to load their own data in source or RDF formats

Optionally we can provide additional upload mechanisms and automation which will integrate with clients' existing workflows to support "business as usual" publication of linked data, including support for near real time data streams



Linked Data Publishing Platform (LDPP) Summary

The platform is customisable and can also host applications running on top of the data.

Our **Linked Data Publishing** service provides setup and support services for the Linked Data Publication Platform, including data modelling and preparation, platform configuration and custom data presentations, as well as migration, testing and ongoing support.

We provide setup and support services for the planning, setup, migration, testing and ongoing support of our **Linked Data Publishing Platform (LDPP)** and the companion **Reference Data Management Platform (RDMP)** offerings.

We also offer training courses for people wishing to develop their own skills in linked data publishing (see our **Cloud Support Training service**).

We offer three variants of the platform for procurement via GCloud.

1. The Flexible platform: this includes all the facilities of the platform together with whatever optional extras the customer desires. A combination of the fully flexible platform and our linked data publishing service will enable the customer to publish linked data from a variety of original data formats, with near real-time updates if necessary, as well as have a range of custom data presentations.

2. The Complete Fixed platform: this includes all the core features described in this document, including fault-tolerance and scalability. When combined with our linked data publishing service this will support a range of custom data presentations.

3. An entry-level platform for people wishing to start linked data publication. The entry level platform runs on a single dedicated machine, so is neither fault-tolerant nor scalable, and will need to be taken out of service during scheduled maintenance. It only provides limited management information.

Our hosting includes support during UK business hours.

Please note that we are continuing to invest significant development effort into our platform to enable it to handle significantly larger volumes of data, deal with more complex queries more quickly, and to enable new functionality. The overall architecture and component structure described in this section is therefore under constant revision.

Platform components

The platform consists of a Data Management Service (DMS) which in turn manages one or more, customizable, publication environments.

Data Management Service (DMS)

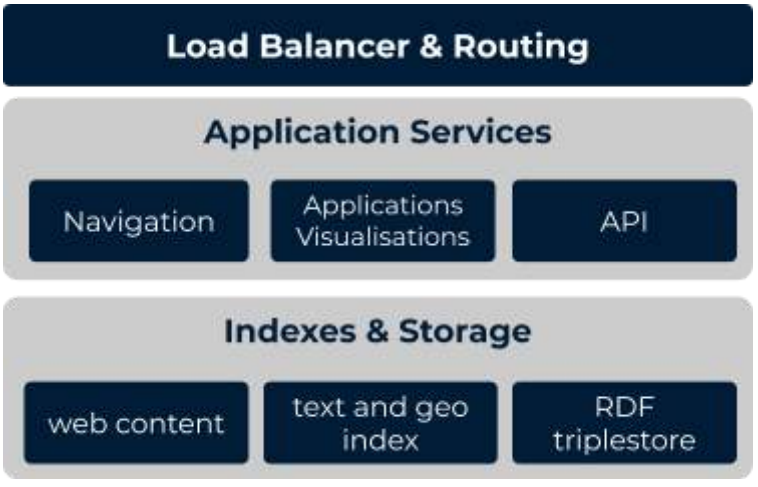
The Data Management Service (DMS) handles all upload and preparation of data for publication, as well as the publication process itself and the management of the publication environments. It supports:

- **Data upload.** We support a broad set of data upload mechanisms including manual upload through web forms, automated fetch of data updates and configurable web service endpoints to integrate into your own data flows.
- **Data conversion.** For all of our platform variants we support publication of data provided as RDF. For the fully flexible system we support upload in a range of formats. Typically data is provided in CSV format and converted to RDF using our open source data conversion tool chain. We also support bespoke data conversion utilities for other formats as required.
- **Data publication.** Controlled publication of new datasets and updates to existing sets can be handled manually through our web-based interface or scripted and automated. In the flexible variant our automation tooling supports a wide range of update frequencies from near real time data flows to large scale periodic bulk updates.
- **Monitoring.** The DMS provides a central point for monitoring and management of the publication environments including the ability to dynamically scale the environments to meet extra demand.

Data publication environments

The publication environments host the published data and any associated APIs, visualizations, applications and documentation. In simple cases a single publication environment may be sufficient. However, some customers find they want multiple different environments for different data services and to support different life cycle stages. For example, a private environment for preview and testing of a data service separate from the public production environment.

The environments themselves are highly customizable. We use a standard open source and standards compliant component stack that we can configure to the needs of different data services. A typical configuration is shown below.



Linked Data Publishing Platform (LDPP) Typical publication environment

RDF Triple Store and SPARQL 1.1 Engine

Our platform is based on Apache Jena, including TDB and Fuseki. This includes the ARQ query engine, which passes the complete SPARQL 1.1 test suite for query, update and protocol.

In addition, the engine can combine free text search with SPARQL queries.

Text and geo indexes

Text search indexing is provided by Apache lucene which can be accessed via an API layer or directly within Sparql queries. We can also provide the option of indexing of geospatial information using either Apache Lucene Spatial or PostGIS.

API Engine

We provide a customizable API to allow developers to access your data in a variety of formats with no need to learn the intricacies of RDF or Sparql. We provide and use a number of open source API tools for this. Our most widely reused API tool (Elda) supports the Linked Data API specification - commissioned by the Cabinet Office and co-developed by Epimorphics. We also offer complementary tool chains to support large scale streaming and batch queries. We can select, configure and tune the API solution to the specific needs of the data service.

Web framework and applications

Our publication environments typically include a landing page to enable users to discover, locate and understand the data together with customizable applications. We typically provide exploration and visualisation tools tailored to the data service, including support for embeddable web “widgets” when required. Tailoring of these facilities is done through our **Linked Data Publishing** support service.

Runtime Platform

The runtime platform can be deployed within several cloud service providers and on a client’s own infrastructure. In this document we assume that the deployment will be within AWS.

It achieves scalability and fault-tolerance by having several identical replicas across different AWS availability zones. Data is kept with the EU for data protection jurisdiction. We can ensure that the data is hosted within the UK if desired.

The replicas are a combination of application services and a local copy of the SPARQL database and any associated indexes.

An Amazon load balancer tracks active nodes and routes traffic based on current load and availability of service machines. The number is adjustable through the DMS to meet the expected load on the system and desired responsiveness within the available budget.

Entry level platform

For the entry level system the runtime platform is limited to a single machine (there is no replication and no load balancing). There is no direct access to the logs of incoming requests. Apart from this the details are the same as those described under “Runtime platform” above.

Optional Additions

Existing workflow integration

Optionally we can can provide additional upload mechanisms which will integrate with a client's existing workflows to support “business as usual” publication of linked data. For example we have integrated publication into existing telemetry process for the Environment Agency to support near-real time data APIs for flood warnings, river levels and rainfall and tide gauge data. This integration builds upon the automation capabilities provided by the Data Management Service.

URI Reverse Proxy Service

Some clients wish to use their data domain for multiple data services. Typically each data service will be hosted on separate servers and may be run by a variety of organizations or providers.

To support this we can also provide a **reverse proxy service**. All requests to the domain would be routed through the proxy service which would *forward* each request to an appropriate backend service, passing the response back. In this way the data is visible to users through a single domain and they are insulated from the details of, or changes in, the backend service providers.

Service details

As a hosted service, our platform is currently accredited to store and process IL0 information only. We are in the process of seeking enhanced accreditation for the future.

All data loaded onto the platform is backed up at the time the data is loaded, so the backup is always an accurate reflection of the data in the system. The replicated nature of the platform means that a hardware failure will not cause data from the running system to be lost. In the event of catastrophic infrastructure failure which takes out all the replicated instances the data will be restored from backup as quickly as possible.

As the system is fully replicated routine maintenance can be carried out without taking the system off-line; there is no need for scheduled maintenance windows when the system is out of service.

We aim for the availability of the system to be 100%. Details of our support services are given in the next section.

We do not offer a trial service though we offer an entry level offering for fewer than 10M triples – see our separate pricing document for details.

Support

Our hosting support for the full system includes all regular maintenance, monitoring and backups. We will provide reports to the clients on the usage of the system – the precise details of the data reported will be agreed with the client during the setup phase. We also provide an incident reporting service. The basic service is available during normal business hours (09.00 – 17.30 Mondays –

Fridays, excluding public holidays). We provide an email address for incident reporting and will respond to any notification within 4 hours. If an incident results in loss of service we will restore the service within 1 business day; in all other cases we will use reasonable efforts to resolve the incident as quickly as possible.

The replicated nature of our architecture is such that we need not take the system down to perform regular maintenance and system updates. The production system we run for the Environment Agency hosts four live services (Bathing Water quality, Flood warnings and river levels, Water Quality Archive, electronic Public Registers) which have averaged 99.998% availability for the last year.
For the entry-level system, running on a single dedicated machine, we will still provide an email address for incident reporting and will respond to any notification within 4 hours; however, if an incident results in a loss of service we will use reasonable efforts to restore the service as quickly as possible, but will not offer a guarantee we will restore the service within 1 business day.

Use of Open Source software

Our platform is based on open source software, notably

- Apache Jena, including ARQ, TDB and Fuseki
- Apache httpd Web Server
- ELDA, Epimorphics open source implementation of the Linked Data API
- Apache Tomcat
- Apache Lucene
- Kubernetes

Compliance with Open Standards

As a linked-data company we have a passion for open standards. We've been key in designing and developing many of the web standards and vocabularies around linked data. For more information on our open standards work see our website at: www.epimorphics.com. Specifically, Linked data depends crucially on the correct implementation of the relevant open standards. Our platform fully complies with all the relevant standards, notably:

- RDF syntaxes: RDF/XML, Turtle, N-Triples
- RDF 1.1 Turtle
- SPARQL 1.1 Query
- SPARQL 1.1 result set formats (XML, JSON, CSV, TSV)
- SPARQL 1.1 Update

Onboarding and Offboarding

We offer support in getting started with our dedicated setup and support service under Cloud Support. We provide documentation and guides to some of our tools through our website (www.epimorphics.com) and our code documentation on Github (github.com/epimorphics). In building skills, awareness and supporting the design of data we offer a number of training options.

If no customisation of the web interfaces etc. is required, then we will provide the client access to the data management system so they can have data loaded onto and published by the platform

within 5 business days after contracts have been signed. We can provide expedited on-boarding at extra cost if desired.

No user data is collected by the system – the only data stored on the publishing platform is data supplied by the client. On termination of the contract all client data will be securely deleted. During the life of the contract clients can request access to a copy of all the data stored on the system. We have also supported public sector clients in transitioning the technology to their own infrastructure.

Sustainability

As a primarily cloud based web technology provider our largest energy usage is that of our cloud hosting and processing: We primarily use AWS, all our AWS provision is within their carbon-neutral regions (primarily EU Ireland). AWS achieve this carbon-neutral position through offset by renewables into the same grid (through direct energy production and the purchase or retirement of Renewable Energy Credits and Guarantees of Origin). Amazon also use resource utilisation and infrastructure efficiency to reduce the overall energy consumption beyond the carbon neutral commitments.

Our other largest external IT infrastructure use is through our collaboration, documentation and email infrastructure. For this we use GSuite (Google). Google matches 100% of the energy consumed with renewable energy and maintains a commitment to carbon neutrality. In addition, Google look to reduce the overhead energy usage alongside their renewable energy commitments.

Or residual energy usage is on-site, for this we use an energy supplier that enables us to only use energy from renewable sources. Our current supplier, Good Energy validates its claim of supplying 100% renewably generated electricity through the Renewable Energy Guarantee of Origin certificates (REGOs) scheme. It also stated that on its main tariff it retires Renewables Obligation Certificates (ROCs) at an equivalent economic value of 5% above statutory compliance levels that apply to all electricity suppliers In addition, Good Energy invests in renewable energy production directly and through providing support to over 800 independent generators in the UK. Good Energy's 100% renewable electricity promise is independently assured by SGS.

Our other Services

We have a complementary **Reference Data Management Platform (RDMP)** offering. This is a solution that builds upon the open source registry software that we developed. The platform and software is trusted and used internationally by organisations needing to manage their reference data. It provides services to create and manage reference data such as controlled, authoritative lists of identifiers as URIs. This supports good data governance, data standards, data collaboration and data use.

We also provide GCloud **Cloud Support** services in support of all our **Cloud Software** offerings.

Linked Data Publishing Support

We also provide support for all aspects of Linked Data modelling, publication and presentation in support of our **Linked Data Publishing Platform (LDPP)** and **Reference Data Management Platform (RDMP)** offerings: this can be procured via our **Linked Data Publishing Support** service.

Linked Data Training

We provide targeted training in support of our **Linked Data Publishing Platform (LDPP)** and **Reference Data Management Platform (RDMP)** offerings through a set of courses. These cover the fundamental principles of linked data and the practical development and use of our linked data and semantic technology solutions. Our courses have been developed based on our extensive experience helping public sector organisations publish open data.