Environment and Energy

Data Service as a Platform

Scaling up: digital data services for the social sciences

The challenge

The arrival of 'big data' has changed social scientists' expectations, and brought technological and infrastructure challenges for data services and repositories. The UK Data Archive (UKDA), lead organisation of the UK Data Service (UKDS), now needs to ingest sizable streams of real-time data, and enable exploration and linkage of a variety of data assets.

For us, this is nothing new. UKDA has been operating since 1967, when 'data storage technology' meant punch cards. We've adapted to magnetic tape, floppy disks and modern online databases – and users have moved from examining printed statistical tables to downloading survey data files, and to exploring and visualising over the web.

In the era of big data, we need to completely review repository architecture and infrastructure – and by focusing first on the field of energy research, we've been able to plan for the challenges we face.

> We're developing an infrastructure to manage new and novel forms of data. As we diversify our holdings to include high-volume energy data, we can also maximise the value of our existing social sciences data portfolio by using the increased computational and storage capabilities of platforms like Apache Hadoop.

Our solution: a modern infrastructure

Social science data archives across the world have long-standing challenges to their existing 'study'-driven repository infrastructure: data management silos, proliferation of different user-facing access points, and difficulties associated with varying levels of discovery and access.

The architecture of a data repository reflects functions and data pipelines defined through standard lifecycle approaches and through reference models like the OAIS (see Fig 1). By infrastructure, we mean the skills, workflows and technology we use to implement the architecture to support our work.

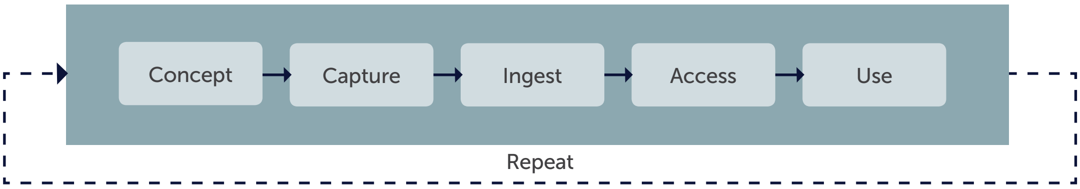


Figure 1: The Capture-Ingest-Access curation pipeline

The architecture will remain familiar for 'big data', but to deal with practical issues of scale, speed of ingest, and the variety of new data sources, archives like ours will have to bring new elements into their infrastructure.

We're developing an infrastructure to manage new and novel forms of data. As we diversify our holdings to include high-volume energy data, we can also maximise the value of our existing social sciences data portfolio by using the increased computational and storage capabilities of platforms like Apache Hadoop. Hadoop sits at the core of our new infrastructure, currently known as Data Service as a Platform (DSaaP).

UK Data Service

From October 2017, new core technical work is allowing us to update our current data service to the new platform. We expect to deliver a proof-of-concept DSaaP infrastructure by the end of 2017. From 2018, we'll integrate extra Hadoop components and other relevant open-source tools, with the aim of rolling out a production-ready unified data infrastructure in late 2019.

Through knowledge exchange and collaboration in energy research between 2015 and 2017, we've built up an excellent partnership with the UCL Energy Institute, and secured significant funding from the EPSRC to develop our Smart Meter Energy Research Portal (SMRP).

New technology like Apache Hadoop and new data paradigms like NoSQL (non-relational databases, which don't fall into a neat grid) mean we can now acquire and process new forms of data, such as rapid streams of sensor data and smart meter data. These technologies also give us opportunities to maximise the value of our existing social science data holdings by restructuring them into the Resource Description Framework (RDF) format, a standard data interchange model, allowing richer description and more granular operations, including data linkage.

For a trusted digital repository, like the UKDA, any solution must:

- be open-source
- be enterprise-ready with the ability to scale up and scale out to petabytes of storage
- have a hardened security infrastructure to deal with authentication (who you are), authorisation (what you're allowed to do) and encryption
- have mechanisms to implement a mature data governance framework

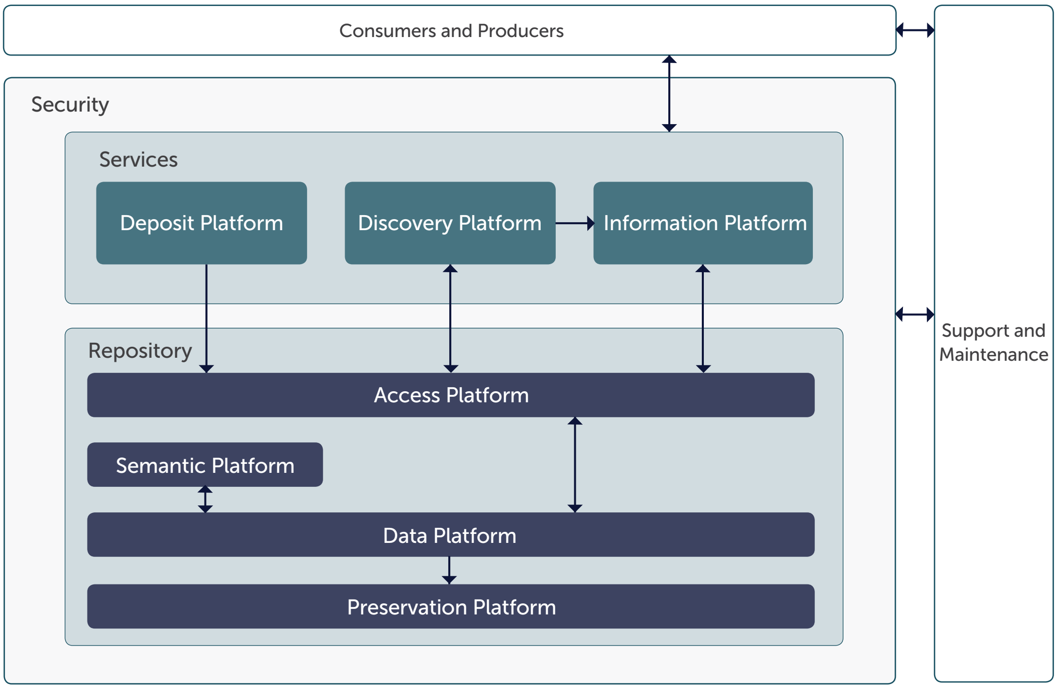
We have implemented an Open Data Platform Initiative (OPDI) Apache Hadoop Distribution, Hortonworks Data Platform (HDP). This 100% open source ODPI Core means we have a stable Apache Hadoop runtime environment and operations, use reference implementations architectures form our implementation and can use common test suites. Our partnership with Hortonworks, a leader in open-source Hadoop-based big data technologies, means we have been able to upskill and support the Service's Big Data Team in delivering our (ODPi Apache Hadoop Core) Data Services as a Platform (DSaaP) solution.

DSaaP offers a holistic platform for managing the Capture-Ingest-Access pipeline while also transforming and modernising it to deal with new forms of data at scale. It also offers us a flexible and cost effective infrastructure that links data on physical servers and cloud-based storage. This 'hybrid model' allows tight security but also gives users a seamless way of accessing the data they need.

Our hybrid solution uses Amazon Web Services as a cloud provider, allowing us to perform on-demand scaling for our search infrastructure and web traffic – and on-site infrastructure to maintain tight governance controls on our local data stores.

The Data Service as a Platform (DSaaP)

DSaaP uses the core HDP product which is specified and implemented as seven discrete but interconnected 'platforms' that handle the entire data lifecycle.



Each discrete platform implements a particular set of operations, known as a microservice. Each can be managed independently, has an isolated security boundary and makes software development more straightforward, as individual components can be created and deployed rapidly and with little disruption

We have adopted many of the tools in the hierarchical data format to manage and orchestrate the data pipeline. Originally devised by the National Security Agency (NSA) in the USA and released as open-source tools, HDF provides full logging and provenance of data as it is processed.

Security considerations must meet our spectrum of data access from Open, through Safeguarded to Controlled data. We operate on the basis of '5 Safes' – safe data, safe projects, safe people, safe settings, and safe outputs, and will use a 'containerisation' approach such as Docker to offer users fully brokered access to data and the tools they need for their research. In other words, by using '5 Safes at scale', users gain access to a non-persistent 'virtual research environment', and storing the data anywhere permanently is avoided.

Overall function of the platforms

The broad functions and dataflows supported by the various platforms within the security boundary will be familiar to most repository managers.

Data producers (depositors) have their data pushed (or pulled) into the repository through a data deposit platform. Deposit is mediated through the access platform to identify/manage the depositor, and to identify the access criteria associated with the data. The deposited data are managed within the data platform, where they are enriched and harmonised for use in the information platform. This harmonisation process is standardised through the application of rules managed in the semantic platform. All actions pertaining to the data and its metadata are defined by and recorded through the preservation platform – the auditable repository workflows which underpin good curation and preservation practice. Enrichment includes the provision of more granular metadata (to support data discovery) and access criteria (to support sub-study level access control), which are managed through the access platform and deployed in the discovery and information platforms.

All potential data consumers can explore metadata by searching and browsing the metadata in the discovery platform. Data consumers who meet the criteria – through research approval, training etc. – may progress to the information platform and can make requests for data which are mediated by the access platform. In some cases, data consumers may be limited to linking and querying against variables and receiving aggregated results. The linkages and queries which are permitted may be based on an underlying set of disclosure risk evaluations about the data to be linked. Other data consumers may have less restricted access to more granular data. All linking and querying happens within the information platform and outputs from this environment may be subject to statistical disclosure control.

Because of the need to offer data linkage and other data manipulation services which the data consumer cannot deploy locally, and/or because the data themselves present a disclosure risk, there is no option to simply approve access and hand the data over to the user in its entirety. From both the repository manager and the data consumer perspectives, the key difference from traditional repository functions is the repository's move away from purely mediating access towards mediation of data use through the information platform.

Pre-defined data products will continue to be made available via the discovery platform, but bespoke data products – such as those generated through data linkage, and other data manipulation services which the data consumer cannot deploy locally, and/or because the data themselves present a disclosure risk – will only be available through the information platform.

While the architecture may be familiar, from the repository manager perspective there is a need for new infrastructure to support more granular access to data for both traditional microdata and new and novel data, and to allow new and novel research. The DSaaP infrastructure supports machine actionable quality assurance and other curation actions at a scale that would not be practical manually. Compared to traditional 'study-driven' paradigms (where a single set of access criteria are applied across the all of the data/metadata in a digital object) the DSaaP enables support for more detailed structural metadata and for the application of access criteria at these more granular levels to data sources pre- and post-linkage.

New staffing skills

Repositories will need to enhance their traditional digital data repository skills to meet the need of this new data infrastructure technology.

New roles for archives include:

- data scientists who embrace the scientific method and can work on large datasets
- data engineers and data software developers who understand architecture, infrastructure and distributed programming
- data solutions architect, data platform administrator, full-stack developer, and designer

Summary of platforms and various tools

Services

Deposit platform

This is responsible for handling ingest of data from across a variety of sources, periodicities and formats.

Technologies: open source Apache products such as NiFi, Kafka, Flume and Spark, and open source Kylo, built on Spark and Nifi. We use Nifi throughout the various platforms to orchestrate the flow of data, availing of its data provenance mechanisms. Our streaming framework combines Flume, Kafka and Spark Streaming to provide a fault tolerant ingest platform.

Discovery platform

On this platform, all data and metadata is discoverable, the level of details is dependent on access permissions using a simple, approachable interface for users who want to conduct simple discovery and visualisation on metadata and aggregated data.

Technologies: Elastic search to allow fast search/discovery over our data stores. Elasticsearch indices are built from our graph data on initial ingest and update, maintaining a nearly real time discovery portal. Kibana, sits on top of the Elastic stack, to provide a consistent user friendly interface for discovery and visualization. Over time custom GUIs will be developed for a more user-friendly experience. A custom plugin for Kibana allows users to explore our DDI graph – a user friendly GUI generating powerful SPARQL queries.

Information platform

This is used for performing more sophisticated and potentially disclosive analysis, brokered through the Access Platform.

Technologies: The HDP Kerberized visualisation component Apache Zeppelin gives technologically sophisticated researchers access to richer data manipulation and querying using Spark2, utilising their syntax of choice – Scala, Java, R or Python. In addition, RESTful APIs will allow data to be consumed by application and machine agents.

Repository

The data platform is where all data and metadata enrichment and harmonisation activities are undertaken to make the data usable within the Information Platform. In combination with the Semantic Platform, we can convert data, ingested in a variety of formats, e.g., SPSS, CSV, TAB and DDI, to RDF triples to form our interconnected graph of the social science universe. Iteratively, as we build ever richer and more harmonized metadata, we provide a more powerful resource for our researcher community

Technologies: Apache HBase and Apache Parquet are utilised for storage of this graph. In combination with the DDI Alliance we are evolving a DDI schema to structure our graph.

Semantic platform

This is responsible for managing ontologies and enabling data enrichment operations performed in the data platform, by human agents or by machine learning algorithms. Ontologies include, thesauri, statistical classification schemes and geospatial structures enabling large-scale geo-linking.

Technologies: Protégé for ontology management and publication. State-of-the-art concept (topic) classification and fall-back manual concept allocation will be combined to enrich variables and categories.

Access platform

This holds metadata related to data producers (depositors) and is responsible for brokering requests between users (human or machine) and data storage. It is a critical bridging component, or gateway, that makes sure access requests are properly mediated and auditable. Our UKDS access model makes sure the licence conditions defined for any digital object will drive the access conditions in a structural and machine-actionable way.

Technologies: currently MS Active Directory is used as the domain controller for our Kerberized Hadoop environment and auditing and governance components of both HDP and HDF allow us to maintain comprehensive data lineage.

Preservation platform

This records relevant actions and version information from all of these workflows as the data travels through the repository pipeline. These support auditable business processes and provide the full provenance of data and metadata management actions necessary to support active long term preservation.

See our case studies on:

- Using smart meter data to enable energy demand research
- Research with household energy data at scale
- Researching the thermal character of UK dwellings
- Amazon Web Services
- Horton Works

Authors:

Darren Bell, Hervé L'Hours, Deirdre Lungley, Nathan Cunningham and Louise Corti, UK Data Service