Case Study:

London Datastore

Type: Website

Organisation(s): Greater London Authority

Tags: open data, process, metadata, standards

LONDON DATASTORE

The London Datastore is a repository for statistical data, metadata, commentary and visualisations on the city region from a wide range of sources, including central government, the Greater London Authority, local authorities, and utility companies. The Datastore includes both an open data platform and a closed platform for secure data sharing.

The Datastore provides a public-facing dashboard with an overview of key information on public services, transport, the environment, health, housing, and demography. Datasets containing underlying data in 18 categories are included separately. There is also a secure private data sharing platform which has become an increasingly important part of London Datastore's work.

Development and web hosting of the London Datastore is provided by DataPress whilst the GLA Datastore team manage strategy and content. Meanwhile, the datastore's API allows file download, queries to tables and programatic updates. The website currently hosts over 6,000 datasets covering a wider range of areas and attracting around 60,000 unique users a month as of 2019.

Background

The London Datastore was established in 2010 by the Greater London Authority, primarily to allow for transparency and scrutiny of the Mayor and the GLA. In the context of the MPs' expenses scandal there was a strong drive towards transparency, which was backed by then London Mayor, Boris Johnson.

At the time, the GLA was also receiving a large volume of calls asking for data to be released, and it took a lot of time to compile bespoke Excel spreadsheets in response to each request. It was hoped that releasing data openly through the London Datastore would reduce the volume of calls, freeing staff up for analytical tasks.

The Datastore initially included 200 data sets, each of which contained multiple aggregated spreadsheets compiling data from across the London boroughs. For example, this included data on GLA budgets, governance structures in London, air quality, and traffic collision data.

Initially, datasets on London that were not directly related to the transparency agenda made up a small proportion of the Datastore's releases. However, it transpired that there was greater demand for data about London. As such, the Datastore has since expanded significantly in recognition of the broader potential of open data to improve governance and drive innovation.

Today, the platform acts as a repository for data sharing between numerous organisations throughout Greater London, including local authorities, emergency services, and third sector organisations.

Budget

The initial budget for the setup of the Datastore was less than £100,000. Even today, the Datastore only has two permanent employees, and the annual fee paid to DataPress provides good value as development costs are split across multiple cities. However, a broad range of analysts across the City Intelligence Unit and the wider GLA also devote a lot of time to producing and curating data published on the DataStore and creating bespoke visualisations. This makes it difficult to define the overall budget, but also highlights how integral the Datastore has become to core workloads.

The GLA is considerating options for the future development of the DataStore, based on the recent ODI discovery work. Whilst the Covid-19 pandemic has led to new financial pressures, successful recovery will require efficient use of and sharing of data which the DataStore will be at the heart of.

London Datastore Rebuild (2014)

In 2014, the London Datastore underwent a major overhaul. This involved increasing the scope of the programme significantly; and finding ways to make the Datastore's ongoing expansion in terms of datasets and publishing organisations smoother.

One priority was to make the Datastore more searchable. Search functionality of the original DataStore was highlighted as being relatively poor, and was particularly hard for users who were not subject matter experts to find what they needed. Adding an improved search engine and better search filters helped overcome this problem.

Another identified issue was that there were only three GLA employees with permission to publish on the Datastore. Publishing was also a very manual process, creating a major constraint as the Datastore's capacity expanded.

At this point, the Datastore introduced CKAN - on the recommendation of one of the partners at the European project iCity - to automate and streamline this process. CKAN, an open source data portal software, allows much better management of publishing organisations, each of which can have one superuser with administrator rights.

The Datastore team put out an open competition for a customised version of CKAN. DataPress, who provide a more flexible cloud instance of CKAN, were successful. DataPress then designed the basic architecture of the current Datastore and still hosts the website.

Secure data sharing: moving beyond open data

As of 2018, the Datastore has begun to move beyond merely open data, towards facilitating access to data across the data spectrum. A new secure level is now a key part of the Datastore's functionality, containing commercial and personal data that it would not be possible to publish openly. Partners range from London boroughs and TfL, to utility companies, universities, and voluntary sector organisations.

There have now been almost as many 'closed' datasets shared since 2018 as there have been open datasets published since the London Datastore was launched.

This platform allows easier cooperation on a whole range of operational areas and is much more efficient than ad-hoc sharing using e-mail or file transfer sites that don't include metadata with the data. For instance, the GLA can share data sourced from the National Pupil Database with individual London boroughs to help predict numbers of schools places.

This makes long term investment planning for schools building programmes much easier to manage.

Another example of secure data sharing is the case of building development. The London Datastore facilitates the sharing of plans for new residential and commercial developments between the GLA, TfL, and utilities companies. This helps ensure more joined-up thinking on infrastructure requirements and investment opportunities.

London Datastore metadata register

A key development that the London Datastore is now working on is building a central register of metadata for all key London datasets, both open and closed. This is not possible in certain instances where even the metadata is sensitive, but in most cases it is achievable. London's highly decentralised makeup means that it is not possible to create a central data lake of the actual data. However, signposting data held by public authorities, universities, private companies and the voluntary sector will greatly aid collaborations on data-led projects.

The role of the DataStore for Internet of Things/sensor data

The European 'Sharing Cities' programme, championed in London by Mayor Sadiq Khan, is an initiative that aims to find ways to connect data held by city regions across Europe to explore future uses of data and Internet of Things technology. The Royal Borough of Greenwich is acting as a demonstrator district for this project and is aiming to implement several transformation projects to improve neighbourhoods and digital infrastructure.

For example, in one estate in Greenwich, heat pumps have been installed as an alternative to gas heating. Smart city sensors have been installed to monitor the success of the project and to understand the benefits of heat pumps, as well as the potential benefits of additional solar plus storage technology. Building on the idea of the DataStore as a central register, details and access to this information is shared on the London Datastore.

Another example is electric vehicle charging. The London Datastore has been working with LOTI (London Office of Technology and Innovation) and London boroughs to compile a single dataset for electric charging points in Greater London. Charging station APIs have been shared to help understand charging profiles and identify where usage is particularly high. This helps charger providers to understand where to install additional charging points and helps the National Grid to understand where it needs to reinforce local energy infrastructure to adapt to higher peak demand.

Important considerations

Content and quality

The metadata is generally of high quality, with detailed explanations of where each data set originates and what it shows. There is a consistent template for metadata, with the data publisher in the subheading, followed by date created, date of last update, a summary, and notes on modifications and points to be aware of.

The most populated data set topic areas are demographics, democratic transparency, environment, and employment and skills**.**

The Datastore data has given rise to spinoff analytics tools to allow more interactivity. The London Infrastructure Map, which includes current infrastructure and planned energy, water, and transport infrastructure projects, is accessible through the Datastore and is built on Datastore data. Another useful feature is London Area Profiles, which allows a dynamic view

of different districts of the city with numerous filters and layers available to get comprehensive insights into their socio-economic, demographic, health, and education statistics.

A newer feature is the built-in analytics tools which provide an additional layer of interactivity, such as a population yield calculator for housing developments.

Challenge-based approach to data publication

In the past, the London Datastore's publication strategy was relatively ad hoc and had a supply-side focus, on the assumption that opening up public data was a public good in itself. However this meant that the GLA were unable to measure the value of the data they published and had limited engagement with the users of the data.

In order to make data publishing more impactful, the ODI recommended in their discovery report that the GLA focus on a challenge-based approach to data publication. This means supporting the delivery of a specific policy (such as supporting the improvement of London's highstreets or reaching net zero carbon emissions) by identifying key stakeholders, compiling a variety of data sources to meet their data needs and delivering this to them as service through the Datastore.

A simple approach to this is to tag existing Datastore datasets under a theme, thus creating a collection of relevant data, such as the page on COVID-19. More ambitious examples include gathering new datasets or building tools on top of the data (such as interactive web maps) to improve usability for non-technical users, such as the Cultural Infrastructure Map.

Data sign-off

The London Datastore's approach to sign-off is to make privacy checks the responsibility of the respective publishing organisation. However, the team does provide advice if any prospective publisher wants to publish a new dataset and is concerned that it might be possible to triangulate individuals in anonymised datasets.

The Datastore has been using the Information Sharing Gateway to create simple privacy impact assessments, and a programme is being developed for a more detailed one in a new ISG module.

DataPress

DataPress covers many different functions at once, including searching, storing, and blogging. The software has also improved significantly over time, supporting new functionalities such as authentification of users.

Whilst DataPress has served the London Datastore well, the Datastore team are now considering a broad range of options for further development. A key priority for the future is to be able to create more seamless integration between different modules, for instance, GIS, IoT, data catalogue, data visualisation and data sharing agreements.

DataPress may be one of a group of software applications supporting the Datastore with separate, bespoke systems connected via open APIs.

This new approach will allow new functionalities, including better integration of the Datastore itself with the Information Sharing Gateway (ISG), reducing manual workload. There is a need for more interactivity, for instance by allowing the ISG to be called up within the Datastore.

Usage

As with other data stores, the London Datastore is reluctant to draw too many conclusions from usage data because there is no way to track how viewed and downloaded data is later used.

However, anecdotal evidence suggests the most popular datasets include local area profiles, which given an overview of the health, education, planning, and demographic profile of each London neighbourhood. Another consistently popular dataset is the GLA population projection data, which is relied on for planning processes by many organisations.

The London Rents Map, which helps private tenants to quickly gain an overview of which areas of London they can afford to live in, is also very popular. This tool is based on Land Registry price data and compiles data that are not available elsewhere.

Strategies for gaining credibility and support

In the initial stages of the London Datastore, working to gain influence and visibility was more explicit. This involved a strong push for organisations and departments to publish data in the Datastore.

However, since then, the most effective strategy has been to rely on the team's own expertise. The London Datastore has gained a reputation within Greater London, primarily because key stakeholders know the platform is an effective and efficient way to share data with a wider audience.

Now policy teams know that they can approach the London Datastore with any ideas about new datasets to publish, and know that they will get a positive reception. For instance, the sharing of the data behind the Culture Infrastructure Map was designed and published within just two days, following collaboration with specialist local groups in the cultural sector to gather the data in the first place.

Blockers and challenges

The GLA has faced several challenges along the way. This has included finding a common cause across such a large project with so many partners. This has made 'community building' a particularly important concern in the project's future management.

Borough-level differences in open data publication strategies have also sometimes impeded comprehensive linking of data and have slowed down the alignment of data held at that level. This includes the need to manage overlapping datasets.

Outstanding issues are also being mitigated by managing data from various partners using High-Level Agreements. The GLA is also building standardised privacy impact assessments (see above) and analytics functions to improve consistency and reduce costs.

What can Greater Manchester take from this?

- It is possible to get a wide range of external organisations on board as contributors, once an effective and engaging platform has been set up.This can be achieved by demonstrating the real value of open or shared data to the organisation in question. Property development is a good example of where the network effect of data sharing is very tangible for all those involved.
- Success depends on constant dialogue and engagement with business and other public sector organisations, as well as strong political support.

- It is the responsibility of any datastore team to demonstrate the value of the services they offer to achieve buy-in. It is important to show professional competence and political neutrality when doing so.
- It is important to provide good search and filtering functionality from the beginning to allow seamless navigation.
- It is important to consider the needs and frames of reference of the non-expert user, who might use different terminology to find a certain dataset than a specialist.
- The future of city data is not just around open data but also shared, closed data. Providing a platform that allows easy sharing of more sensitive data between local authorities and other public sector and external organisations has major benefits in improving collaboration and communication.
- Providing a central repository of metadata for all datasets relating to Greater Manchester and published in different datastores could significantly improve user experience. This helps to avoid duplication of publication workload and provides a simple overview for data users with reduced search time.
- A challenge-based approach to data publication can help ensure that data publication is driven by a clear purpose and focuses on the most useful datasets. To make this relevant to all possible data users, it is important to run broad consultations involving key stakeholders around each challenge.
- Large datastores can be built and maintained on a relatively small budget provided they are integrated into existing workflows and capacity.
- When co-ordinating a large project across many external partners, 'community building' may need to be a central part of the project management.

Find out more:

https://medium.com/@SmartLondon/10-years-of-the-london-Data Store-thinking-on-city-data-for-the-next-decade-b634ae62dc3c
https://medium.com/@SmartLondon/london-Data Store-turns-9-today-eec8f59e439b
https://medium.com/@SmartLondon/a-smarter-london-together-listening-exercise-for-a-new-smart-london-plan-51be7d9ca203
https://www.london.gov.uk/what-we-do/business-and-economy/supporting-londons-sectors/smart-london/sharing-cities
https://data.london.gov.uk/blog/the-morning-after-the-night-before-international-recognition-for-the-london-Data Store-2/
https://www.london.gov.uk/sites/default/files/city_data_analytics_programme.pdf
https://localgov.digital/wp-content/uploads/2019/03/GLA-LGDSS-review-London-Datastore-1.pdf
https://theodi.org/article/discovering-the-future-of-the-london-datastore/