Case Study:

Data.gov

Type: Website

Organisation(s): US Government

Tags: open data, process, standards, website

Data.gov is a large catalogue of US government datasets. These datasets cover a wide range of topics, from business and economy to demographics and environment. The information is sourced from federal agencies and local and state governments. The website predominantly provides links to URLs of datasets held by other organisations or to external datasets.

The US government's strategy on open data has been bolstered by a series of federal acts and decrees, which have progressively increased the ambitions of the government's open data strategy.

The size of data.gov makes it a very useful one-stop-shop for data on the United States, but its scale also highlights challenges around standardisation and data quality.

Background

Data.gov was launched in 2009 following Barack Obama's Memorandum of Transparency and Open Government, which he signed on his first day in office. The 2009 Memorandum was signed in a political environment in which open data had become increasingly topical. In 2007, a conference had been held in Sebastopol, California, where thirty open data advocates drew up a list of eight principles of open government data. The Memorandum was also informed by the principles of the Sunlight Foundation and the Open Knowledge Foundation.

The President followed this up with a further executive order in 2013 which aimed to institutionalise effective information management of public data in the USA. Among other things, this executive order highlights how the needs of the end-user must always be kept in mind. It requires that information collection activities 'support downstream information processing and dissemination activities'. The executive order also points to the need to 'maintain internal and external data asset inventories' and to 'clarify information management responsibilities.'

Another key milestone towards open data in the USA was the 2014 Digital Accountability and Transparency Act (DATA), which required all federal agencies to publish standardised spending data. From 2017 this data was available on the US government portal usaspending.gov. This portal provides charts and geospatial visualisations that present tax, spending, and borrowing data in an easily accessible form.

The work of Data.gov has been further modified by the Foundations for Evidence-Based Policymaking Act of 2019, which requires the US federal government to make its data available in an open, machine-readable format, whilst continuing to ensure privacy and security.

All federal agencies are required to create a full inventory of all their data assets, and also a list of all data assets that are public or could be made public in the future. The result of this is

that, at least in theory, there is an 'open by default' approach, whereby only those datasets which contain private or security-sensitive data remain unpublished.

Data.gov is powered by a combination of CKAN and WordPress and all the code behind it is made available publicly on GitHub. The platform is hosted by the U.S. General Services Administration.

Action 20

Action 20 is a plan to create a data standards repository for US federal agency data by December 2020, to help accelerate the adoption of unified data standards nationwide. This initiative is organised by the General Services Administration (GSA) in collaboration with the Office of Management and Budget (OMB). This will be a user-friendly website designed for technical and non-technical users alike, including case studies, guidance, and data governance. It will link together resources.data.gov, which provides policies, tools, and case studies of data usage, with the National Information and Exchange Model, and other Federal and non-Federal agencies.

Important considerations

Content and quality

In total there are over 211,000 datasets on the Data.gov website. The largest category of datasets on the platform by far is from local government, with 20,961 datasets, with climate the second largest at 592, and education the third largest at 463. The most common data formats meanwhile are HTML, at 99,057 datasets, followed by XML at 37,206 and PDF at 25,779. Meanwhile, the most common publishing organisations are the National Oceanic and Atmospheric Administration (75,170 datasets), followed by the Department of the Interior (26,434 datasets) and then NASA (24,874 datasets).

Some of the datasets are very internally complex, indicating that the datastore is even larger than initially appears. For instance, the Hourly Precipitation Data is made up of fourteen unique datasets, including a link to the National Centers for Environmental Landing Page. This landing page in turn provides multiple access options, distribution formats, and extra resources.

The datasets are easily searchable using free text search and filters. The number of views is also visible for each dataset and sub-dataset, as is the star rating of the dataset's openness. Another useful feature is a map tile for each dataset so users can see exactly which geographic region each dataset relates to.

There is a contact function linked to each dataset as well, which allows users to report issues and make suggestions using a contact form. Users can also request new data and suggest new features of the website. The Data.gov team also keeps active on Twitter, giving updates on important new datasets and features appearing on the website.

Data quality

Data.gov is a catalogue that does not generally host datasets but instead relies on the quality of over 200,000 links to external websites. This causes occasional data standards issues, which is part of what the Action 20 initiative is seeking to address. Inevitably, this decentralisation means that some of the links to dataset downloads no longer work as it is impossible to check all of them regularly. This can be seen for instance in the dataset containing numbers of accidental drug-related deaths in the state of Connecticut.

Also, many of the datasets that are held in the Data.gov catalogue are only available for individual states or cities, when in fact it would be useful to have them for larger geographies. Where datasets have slightly different names or content, they have to be manually collated by the data user. For instance, both Maryland and Connecticut have posted data on drug-related deaths, but the Maryland data also include alcohol-related deaths, so they are not directly comparable. It would be useful to have more linkages between datasets, but this is always dependent on priorities set at a local or regional level.

Usage

Whilst it is not possible to rank datasets by viewing numbers, an article from 2017 revealed that the most popular datasets at the time were, in order of popularity:

- The College Scorecard, which ranks universities across the USA according to a range of measures
- Crime figures from the City of Chicago
- The Housing Affordability Data System
- The Consumer Complaint Database

There is a section of the Data.gov website providing case studies on where data has proven crucial for effective government. One example is the emergency response after the 2017 Puerto Rican Hurricane, where in the first instance imperfect address data made it hard to coordinate the response. This has led to a concerted effort to clean existing data through the Puerto Rico Address Data Working Group.

Developers and APIs

The Data.gov catalogue is powered by CKAN and is linked to a CKAN API. However, this only includes metadata about datasets and URLs linking to datasets and to find individual APIs it is necessary to visit individual government agency websites.

The data.gov website also links to challenge.gov, which is a repository of US government problems for which agencies are actively looking for support from external developers. One of these challenges is an Automated Streams Analysis (ASA) system, which would work with live streaming data on emergency events to detect and analyse them automatically. There is prize money offered for each successful bid, which in this case totals $150,000. This helps keep the developer community actively engaged in finding applications for US government data to find solutions to key issues.

Analytics tools based on data.gov data

Many applications have been developed using data provided by the US government, both by government agencies and external developers. Data.gov also publishes a list of web applications that make use of the data available on the website.

One example is the Price Watch function on the food security portal, hosted by the International Food Policy Research Institute. The application displays wholesale prices for key agricultural commodities like rice, wheat, and soybeans, and measures the level of price volatility for each of them, as well as highlighting global hotspots of food security news.

Farmers can also make use of the Crop Trends app provided by Farm Plenty, a software company that offers both apps available to the general public and apps for individual farmers to keep track of their operations. Crop Trends allows farmers to select an agricultural region and see what crops are grown where and how prices are moving for each crop. Both of these applications help provide farmers with the data resources they need to improve their future planning capacity and decide which crops to grow.

Another example is the Alternative Fuel Station Locator, which shows the location of all public electric charging stations, as well as alternative and non-fossil based fuel stations (such as LNG biodiesel) in the USA and Canada. This is a US government-hosted website that uses existing data as well as allowing users to submit the locations of any new fuel stations or charging points.

Blockers and challenges

The Government Accountability Office (GAO) highlighted some of the key barriers to compliance with the DATA Act in its 2017 report. The report highlights the difficulties involved with establishing a unified data governance framework across all of US government operations. Issues were also identified with data quality, with missing and replicated datasets on certain government transactions continuing to cause problems in the establishment of a US government spending portal.

A wider issue for US government data more generally is the very low levels of public trust in the collection and processing of data by the US government. This is according to data published by the Pew Research Centre in 2019. 78% of respondents said they had little understanding of data collected by the government, whilst 76% said they benefitted very little or not at all from US government data. When asked about specific uses of government data, such as data on poorly performing schools, respondents tended to be more supportive. However, these findings still indicate medium-term issues in winning public trust around government data and transparency.

What can Greater Manchester take from this?

- Setting data-related challenges to the developer community and providing prize money could be a way to help kick-start the discovery of possible new uses of public data.
- Applications similar to those using US government data to provide easy overviews on crops and commodity prices could also be utilised in Greater Manchester to provide market insights on key sectors in the city region which could be of high commercial value.
- If individual local authorities and other organisations begin publishing on a GM portal, it is important to find ways of linking similar datasets even where data might be collected in slightly different ways in different localities. This could help prevent replication of work and allow simple comparisons to be drawn.
- There are major public concerns about government uses of data that make a clear communication strategy and a clear demonstration of possible use applications particularly important.
- Even large, nationwide data catalogues can have the functionality for simple feedback and suggestion forms. This could also be very beneficial for Greater Manchester, particularly if there were some built-in analytics functions to help make sense of feedback more cost-effectively.
- The best way to achieve unified data standards in Greater Manchester may be to first understand the standards that are currently in use in the region (as in the US Action 20 plan) and then capture their benefits and pitfalls. This could help to both ensure more effective standards and achieve more widespread buy-in.

Find out more:

https://www.data.gov/meta/

https://www.data.gov/highlights/

https://www.forbes.com/sites/metabrown/2017/04/30/these-are-the-10-most-popular-datasets-on-the-us-government-data-portal/#6fbd82fa748b