Case Study:

Data Mill North

Type: Website, Team

Organisation(s): Leeds City Council, ODI Leeds

Tags: open data, process, metadata, standards

Data Mill North (DMN) is a large datastore run by Leeds City Council, which holds data sets from a total of 64 publishers, including local authorities in the north of England, as well as many private and third sector organisations.

The Data Mill has a collaborative and open approach that has stimulated interesting and creative uses of data. However, as a result, the team recognises that there is a lack of consistency and unity across datasets.

The team is now developing the concept for a new iteration of DMN, 'Data Mill North 2.0', which will further heighten ambitions, aiming to create a more user-friendly and engaging platform with more connections between related data sets.

Background

Data Mill North (DMN) was originally launched as Leeds Data Mill by Leeds City Council in 2014. It was established with a setup grant of setup grant £165,000 from the Cabinet Office's Release of Data Fund.

As Leeds Data Mill, the focus was on data sets exclusively relating to the city of Leeds, but in 2016 it was rebranded to Data Mill North. This built on a recognition that open data was not just about a specific council or city but should have a regional focus. Many datasets held by other organisations, such as by utility companies, are relevant across local authority borders because of commuting and other ties.The change to Data Mill North involved expanding the datastore's remit to include more councils, private companies, and third sector organisations across the north of England. These are not merely in the Leeds city region, but also in the North West, (e.g. Stockport Council), and the North East (e.g. Newcastle City Council).

DMN has been consistently supported by the Chief Executive and Chief Information Officer at Leeds City Council. Council leadership has recognised the power of the Data Mill as a tool to improve responsiveness to FOI requests. It is also part of a broader digital agenda, Smart Leeds, which aims to improve the quality of life in Leeds through technology. This includes the creation of Living lab at White Rose

Park, where the focus will be on testing and trailling technology that addresses city challenges.

The Data Mill team is currently exploring how the publication of data could be better aligned with the work carried out by those managing the Leeds Observatory, a separate team within Leeds City Council, maintained by the Intelligence and Policy Service. The Observatory conducts some similar work to that of the Data Mill team, such as providing overviews and visualisations of nationally available Socio Economic data. Their work covers a number of areas, including deprivation, crime, and education, which is why a more direct link between the two teams is being explored.

Collaboration and Communication

Publishing Organisations

Leeds City Council has successfully encouraged many external organisations to share data with DMN, including West Yorkshire Pension Fund, Yorkshire Water, National Rail, and the University of Leeds. The first private sector organisation to publish data on DMN was Yorkshire Water, which has ambitious data transparency aspirations and wanted an outlet to publish its data sets. The company views itself as a pioneer in driving the open data agenda among water companies. This is based on the aspiration that external developers will find new solutions to pollution and leakage problems, and also that sharing data will challenge publishers to improve data quality. Examples of Yorkshire Water datasets include leakage data, which helps guide infrastructural investment priorities; pollution data; and the water usage survey, which gives insights into consumption trends.

Relationship with ODI Leeds

Leeds City Council are a sponsor of the Open Data Institute (ODI) Leeds node. Through collaboration with ODI Leeds, the Data Mill team has been able to reach out to external organisations, and drive a focus on the potential uses of data sets on DMN. ODI Leeds were also instrumental in building relations with partners like Yorkshire Water.

Cooperation between the various participant organisations is driven by an Open Data Collaboration Group, which is hosted by ODI Leeds and meets up every 6-8 weeks. This helps create better and more secure lines of communication between partners, and enables the sharing of best practice.

Innovation Labs

From the beginning, the team behind Data Mill North has actively sought engagement with the wider open data community through day-long Innovation Labs.

These Innovation Labs are set up when council services or residents approach the team with a novel challenge that requires a data-led solution. A number of developers are present at each of these Labs and the participants aim to co design solutions that can be shared more widely.

One recent case has involved establishing how many ‘grey miles’ staff travel to work in their cars. Other examples include the Age-Friendly Innovation Labs, which have included a session on accessibility that looked at ways to collate existing data and gather further data on disability access.

These often open-ended sessions help to articulate new questions and ensure that existing datasets do not get overlooked. Potential new uses are identified, which help drive the development of DMN.

Some very useful tools have been developed based on past Innovation Labs, such as the Leeds Bin App, which allows residents to enter their postcode and automatically get notifications the day before a bin collection. This has had a measurable impact, leading to a reduction in missed bin collections by 50%.

Another example is the Social Housing Picker, which is maintained by Leeds Homes and allows users to enter their family circumstances and find out which areas of Leeds have the shortest waiting list. This calculation is based on lettings over the last five years. This tool empowers residents to make better decisions and helps families avoid excessive waits for social housing by putting data in an accessible and usable form.

Data Mill North 2.0

Data Mill North 2.0 is a strategy to revamp Data Mill North and include a range of additional features. The Data Mill team is examining how to link data sets to make them more useful for developers and members of the public alike. One proposal is to create schemas that would include a standard format for any related data across different local authorities. This would allow council officers to more easily access data on other services within the council for their own uses.

To take one example of how this would look, Northern Powergrid is in the process of encouraging the adoption of a 'Distribution Future Energy Scenarios' data standard’. This would involve a data product that maps the power network and energy production infrastructure across the north of England.

The Data Mill 2.0 strategy also includes an ambition to publish real time data and a clearer focus on improving the use of geospatial data to allow overlaying different geographies. It involves a renewed focus on high-quality visualisations more generally, since the Data Mill is currently largely made up of raw CSV files.

The strategy also includes looking to other parts of the data spectrum, including closed data. This could enable closer collaboration with universities and other public

sector organisations on more sensitive data sets, by providing a space to share data between organisations.

This new iteration will also include more automation of data set publication with notifications sent automatically to the team when a new data set is introduced.

Important considerations

Use of DataPress

DataPress designed the Data Mill portal and is the current domain and platform provider. Leeds City Council entered into an agreement with DataPress allowing external organisations to each publish up to 30 data sets on the Data Mill for free. Each data set can include many individual spreadsheets, and external data publishers are generally able to stay well within this limit.

Data Mill North is currently planning to move beyond using only DataPress. The current consensus is that the provider could be part of the solution. The team wants to ensure that the chosen solution for Data Mill North 2.0 is flexible enough to suit the requirements of a more sophisticated data platform.

Impacts for the Council and city

When working to increase commitment to open data, the Data Mill North team tries to focus on what works at a practical level, including improving the speed of FOI request handling. This also includes highlighting where potential pitfalls have been averted. Writing good quality reports based on DMN data which provide new insights also demonstrate potential opportunities further down the line.

Moreover, increasing the use of visualisations helps make the data sets more accessible both by non-expert users within the council and beyond, raising the profile of open data and of Leeds as a data city. Data Mill North and the approach to innovation using open data, has helped cement Leeds as an open data city, as indicated through the numerous awards that DMN has won. These include the Nesta New Radicals award, as well as the 2016 ODI Open Data Publisher award.

Data publication process

Decisions on data publication are based on a number of factors and does not follow a set procedure. The initial approach was one of open by default, which still remains an ambition. However, linked to some of the blockers and challenges presented below, practical limitations have necessitated a more targeted approach to releasing data.

The process for data release is predominantly driven by current FOI requests, but also wider discussions on key council priorities, such as housing, health, climate

change and COVID-19. The open data team is in regular contact with council services to see what further data sets might be published on DMN, as well as learning from the best practice of other participating councils.

Content and quality

There are 807 data sets in total on DMN, which split into twelve thematic areas, each of which is well populated. The most populated ('Local Services') includes 140 data sets the next most populated is 'Transport', with 96, and even the least populated one, 'Sport', contains 25 data sets. By far the most common file type is csv (424 data sets), compared to only 34 data sets that have a JSON version.

Meanwhile, despite diversification, Leeds City Council is still by far the largest publisher, with over 300 data sets, including those by DMN and Leeds Observatory (a unit within Leeds City Council). The next largest publisher is Newcastle City Council, with 33 datasets. This is partly reflective of the limited number of datasets each partner organisation can publish.

Data standards vary significantly by organisation and department. Generally, metadata are quite comprehensive, and always includes a publisher, date of creation and the date of last update. However, there are still obvious gaps, with some data sets published without a description.

Where common standards have been established, however, there have been very positive results. For example, the way that business rates are recorded has been unified across Leeds, Bradford, Calderdale, and Stockport. This has allowed a comparison of different authorities, helping councils to understand where there are more empty properties and how much money they are losing as a result.

Usage

Google Analytics data are collected but are not given much weight in decision-making processes. This is because it is difficult to determine how datasets are used once they are viewed and the number of views often does not reflect the full picture of data usage. One view on the Data Mill by a higher-level data user can lead to thousands of views further down the line if it is published or used elsewhere.

However, analytics reveal the most popular data sets are public health funeral data and the houses in multiple occupation register. This may be because of the clear commercial benefits associated with these datasets: funeral data provides leads to companies searching for unclaimed estates; whilst the HMO register is used by letting agents advertising their business to landlords.

Blockers and challenges

One major hurdle for Leeds has been that, unlike other councils, it does not have a central Business Intelligence Team. This means there is not a unified workflow and there are not even common standards for data sets provided by Leeds City Council itself. As a result, the council has published over 300 data sets itself, but there around 150 different sources of these data in the council. This means different reporting, column naming, and metadata standards, which is even more of an issue when considering the large number of publishing organisations that contribute to DMN.

Another challenge for the Data Mill team is getting hold of the data in the first place. For most individuals in the council who collect the data, adding the information to Data Mill North is just one of many responsibilities, and additional to their workloads. Sign-off agreements with heads of service help ensure access to data, but these are not sufficient to secure genuine buy-in and commitment at officer level. This puts the onus on the Data Lab to tidy up data sets and publishing them to DMN, a time-intensive task.

Moreover, there are fears within councils about publishing data at all, especially around the potential for reputational damage if figures are published in areas where the local authority performs less well. There is a fear over the lack of control over the way that data sets are used once they are published, demonstrating that open data and a wider culture of openness are strongly related.

User Experience

There is acknowledgement within DMN that there are some unresolved issues around usability. There are not many visualisations available, so that it is difficult to gain an easy overview of many of the data sets. Also, whilst it is easy to find the right data set with good search functionality, a lack of consistency in how data sets are published has caused ongoing problems.

For instance, some data sets contain a single file for all years, whilst others are split into individual spreadsheets for each year. This can make it difficult to compare similar data sets, even within organisations.

What can Greater Manchester take from this?

- The bigger a datastore becomes, the more bureaucratic it can become. Without the existing groundwork facilitating automated operations, running the datastore can become extremely time-consuming. If these were to be established from the beginning, it would be much easier to quickly scale up the datastore.

- Conversely, a less centralized system with federated control over publishing and standards can lead to a lack of standardisation and usefulness across the board. A balance between curation – including support for making data open – and making the publishing of data easier is a key determinant of success and usage.
- There is a danger that updated a datastore will be seen as just another piece of work by council officers and more of a box-ticking exercise, so that open data is not considered a core function of an organisation. It is important to clearly demonstrate the benefits of open data to council employees to secure active rather than reluctant engagement.
- Opening up datasets also relies on a culture change within any organisation, because open data inevitably demands willingness to receive scrutiny and potential critical receiption if certain figures show the organisation in a negative light.
- It is important to segment out the users of the datastore based on their needs, and providing a separate closed and open data space helps to do that. Also, whilst data specialists appreciate APIs to keep track of dynamic datasets, the general public want data in an accessible and appealing format,
- In terms of wider usage of datasets, numbers of downloads or analytics on website usage is no guarantee that the data is being well used. Effective evaluation work would need to look beyond headline figures available on Google Analytics.
- It is possible to encourage external organisations to devote time and effort to share data sets with a council-led datastore. This is made much easier where there is a separate organisation such as the Open Data Institute to facilitate this collaboration.
- Initiatives like the Innovation Labs help to bring a level of dynamism which is difficult to achieve within any single organisation, offering fresh external perspective and helping bring in talent on board from the wider community.
- It is very important to align metadata and reporting standards as early as possible between different departments and organisations contributing to a local data programme. It is difficult to unpick this later on, and unified standards allow data users to draw parallels and conclusions much more easily.

Find out more:

Data Mill North website

ODI Leeds webpage on Data Mill North 2.0