Open Data Manchester

The Future of Open Data in Greater Manchester

December 2020

Prepared by Open Data Manchester for the Greater Manchester Combined Authority as part of its Local Data Review, in support of the priorities set out in the Local Industrial Strategy. With many thanks to the people who participated in the consultation.

Contents

Table of contents 2

Executive summary 4

Business 4

Academia 5

VCSE 5

Common challenges 5

Shared solutions 6

Context 7

Methodology 8

Responses 9

Understanding open data 10

Current open-data use 12

Socio-economic 12

Environmental 12

Health 12

Mobility 12

Geo-spatial 13

Education 13

Safety and security 13

Regulatory 13

Democracy 13

Crowdsourced data 13

Other identified data – not necessarily open data 14

Applications of data – business sector 14

Geography 16

Applications of data – academic sector 17

Applications of data – VCSE and other 18

Data that should be made available 20

Socio-economic 20

Environmental 20

Health 21

Mobility 21

Geo-spatial 21

Infrastructure 21

Safety and security 21

Democracy 22

Other comments 22

Challenges and concerns when accessing public sector data 23

Technical 23

Quality 25

Consistency 25

Continuity 26

Risk 26

Skills 27

Health data 27

Locally collected, nationally collated data 27

Overcoming challenges and recommendations 28

Foundational recommendations 28

Technical 28

Quality 28

Consistency 29

Skills 29

Health data 29

Locally collected, nationally collated 29

Promoting open data use 29

Datasets 29

Conclusion and next steps 31

Annex 1. Survey Questions 32

Executive summary

In summer 2019, GMCA (Greater Manchester Combined Authority) launched the Local Industrial Strategy to help set long-term priorities for our regional economy, ultimately to ensure good jobs and better public services locally.

This was the first of such local plans in the UK, created after a consultation with local people and reviewed by expert analysts, and it highlights the importance of data – including its use and re-use – to places like Greater Manchester.

Open data was found to be a strategic priority, particularly its potential to help support transport changes in pursuit of a greener society, but more generally, so it can be used by anyone to make life better for everyone. At the time, GMCA said that:

> "The assumption will be that data that can be made available should be made available, unless it is prohibitively expensive or not appropriate (for legal, commercial or security reasons) to do so."

As such, a Local Data Review was proposed, to identify the opportunities and challenges for the use and reuse of local open data. Open Data Manchester (ODM) was asked to help fulfil the region's ambition to understand:

- what open data is being used
- how it is being accessed
- What additional data could or should made available
- the current challenges associated with accessing open data

To do this, ODM hosted six focus groups covering **academia**, **business**, and the **voluntary, community and social enterprise** (VCSE) sector, along with an accompanying survey.

Across all three targeted sectors, the participating organisations use a wide range of open data, covering all aspects of people's lives in Greater Manchester. Whether in academia, business or charities, open data has become crucial to many organisations' work

Business

In the business community, participants ranged from micro-businesses up to large companies with more than 250 employees. Open data-enabled activity within businesses fell across four areas:

- data analytics and modelling services
- internal, data-analytic functions
- compliance services based on open data
- creation of value-added data

The majority polled provide services at a national level, a trend that was represented in the data needs of those organisations.

Academia

Healthcare themes feature heavily, particularly a need to access a broad selection of data related to socio-economic and health issues, in order to further research goals. Although much of this data is restricted, using anonymised or synthetic data could make more data of this kind available as open data. Local open data is also well-used in teaching, allowing students, who tend to live in Greater Manchester, to understand the data and the context from which it is derived.

The use of open data within the academic sector mainly relates to:

- methods
- training
- research

VCSE

The consultation revealed the varied ways that the VCSE sector uses open data, with charities using socio-economic data to:

- identify need
- target resources
- develop Key Performance Indicators (KPIs)

In particular, topographic data and Energy Performance Certificates (EPC) are used in social housing, and there has been implementation of data standards around health and leisure activities, so that opportunities are easily discoverable for residents.

Common challenges

What was highlighted is a need for better quality data, which is consistent and well-maintained to common data standards, and that is supported by good-quality metadata across the 10 boroughs of GM. While there were specific challenges found within each sector, common challenges (explored in more detail later in this report) were raised around:

- Technical formats
- Quality of data
- Consistency of data
- Continuity of supply
- Usage risk

There were also more specific concerns raised relating to:

- Health data
- Data collected locally but collated nationally
- Skills

Shared solutions

Speaking directly to these challenges and concerns, recommendations have been collated using the contributions of participants, some of which are foundational, and some more specific. These should help guide the deliverables, and actions, for GM, so that more, better data can be opened for public use. The foundational recommendations are:

- Open-data infrastructure should be treated as essential infrastructure
- Develop an open-data strategy that is joined up with GM's other strategies
- Design a central repository with good user experience as the first place for GM's open data and ensure it is well maintained
- When publishing open data, create clear documentation and ensure that supporting materials and metadata is included, and kept up to date
- Maintain an up-to-date data catalogue
- Ensure that open licences relevant to the data are explicit
- Create persistent links for datasets to minimise broken data links

Further, more general opportunities have been identified, in relation to:

- Promoting open data use
- Selection of datasets

A process of continuous engagement with the data-reuse community should be developed to ensure a successful open data programme in Greater Manchester.

Context

In the UK, Greater Manchester has the largest digital sector outside the South East of England. Data is a key driver for this, with it often underpinning the development of new and innovative products. Data can help us understand how our cities function, drive process efficiencies and identify solutions for complex problems that we face today. It allows applications that utilise artificial intelligence and machine learning to function, generating new insight and knowledge.

Open data released by Greater Manchester's public sector organisations has the potential to support existing digital businesses, as well as enable the development of new economic activity. This potential is highlighted in recent study done by Deloitte for Transport for London (TfL), showing that open data released by TfL was estimated to:

- directly support 500 jobs in the digital sector
- plus a further 230 jobs in the supply chain
- offering total Gross Value Add (GVA) per year across the supply chain and wider economy of between £12m and £15m
- plus additional savings of between £70m to 90m per year because of greater certainty of arrivals and departures

A recent Capgemini report, *Economic Impact of Open Data - Opportunities for value creation in Europe,* identified that in 2019 the size of the open data market in the European Union (EU) was €184.45bn. It is expected to grow to between €199.51 and €334.20bn by 2025, employing around 1.12m up to 1.97m employees.

As part of GMCA's Local Industrial Strategy published in 2019, it was identified that a Local Data Review would be undertaken to identify and address barriers to making GM's public data openly available for re-use. This consultation is part of that review.

Methodology

Although the focus of the consultation was to understand open data usage and requirements within the business sector, open data is used more broadly and often sectoral distinctions can seem blurred.

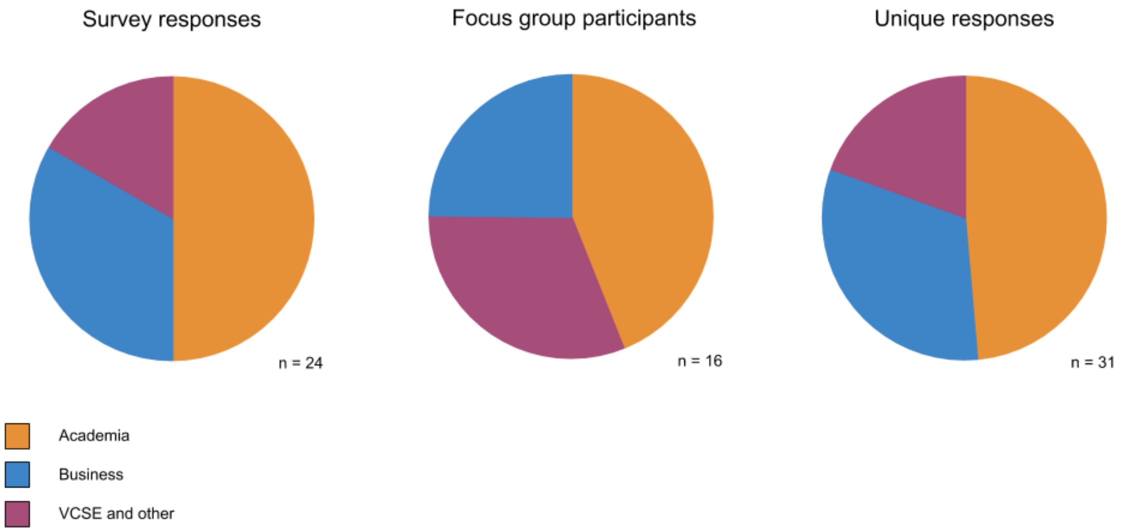
This was evidenced during the consultation, in that four of the businesses involved had interests outside of the business sector, two with academia and two with VCSE. To this end, the consultation was also promoted to the academic and VCSE sectors.

The consultation was split into an online survey and follow-up sectoral focus groups, with the purpose of exploring in more depth the questions that appeared in the survey. The survey was predominantly qualitative and asked 18 questions, with 11 of those replicated in the focus groups. The survey questions can be found in Annex 1.

Due to the constrained timescale of the consultation, participants were found through social media (predominantly Twitter through GMCA Digital @gmcadigital and Open Data Manchester @opendatamcr), direct email and the Open Data Manchester email newsletter. These methods of contact led to the risk of the respondents not being totally representative of the broader open data community.

The survey ran between the 9 and 20 November 2020, with six online sectoral focus groups split between business, academia and VCSE running 24 to 26 November 2020.

Responses



Sector	Survey (full response)	Focus group	Both survey & focus group	Unique responses
Business	8	4	2	10
Academia	12	7	4	15
VCSE and other	4	5	3	6

Partial responses = 37
Full and partial responses will be shared online as a final Annex to this report.

Understanding open data

Open data can be defined in a number of different ways. The Open Knowledge Foundation definition is that:

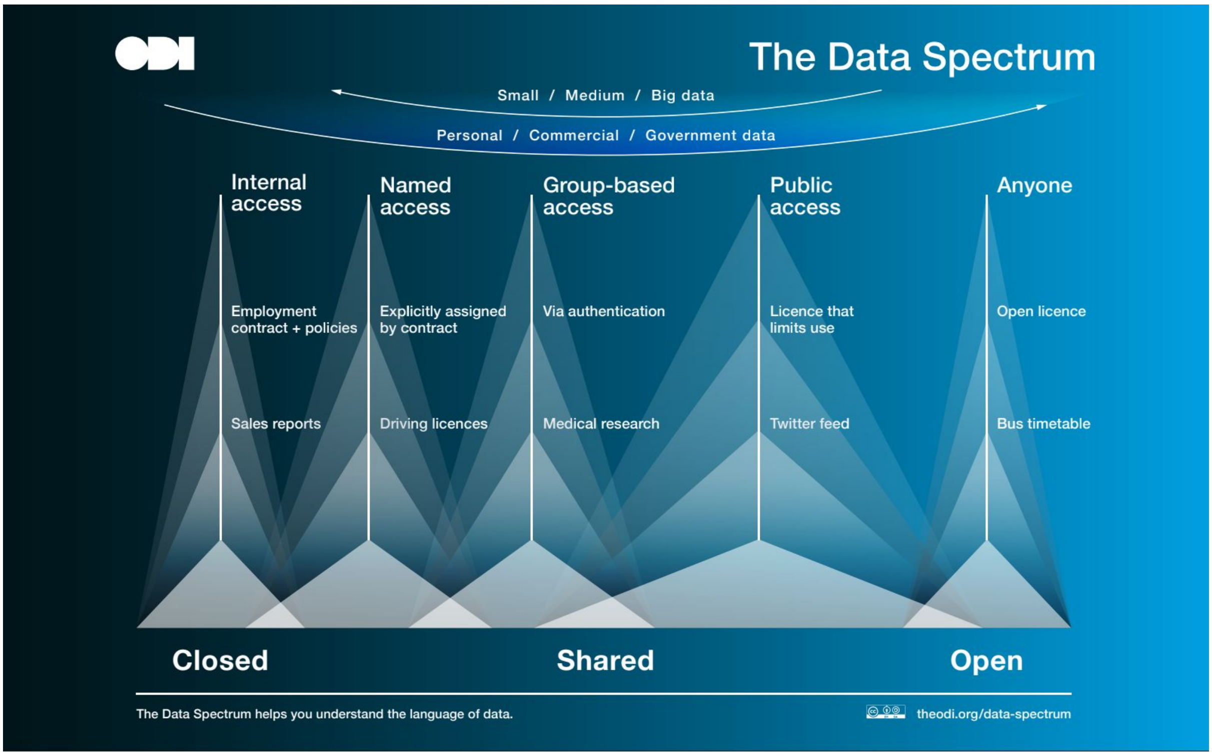
> *"open data is data that can be freely used, shared and built-on by anyone, anywhere, for any purpose."*

The ODI defines it simply as:

> *"data that's available to everyone to access, use and share"*

with the emphasis of *"freely"* removed.

Understanding exactly what organisations perceive as open data is important, as it can help define the approach of future open-data initiatives. The question of what participants considered to be open data was raised in the focus groups. The ODI data spectrum is useful here as it shows the wealth of data that could be made available as open data, as well as what data can be considered closed or is in need of processing in order to make it available as open data.



The ODI Data Spectrum

Respondents saw that open data was a public good, creating value for data publishers, as well as the organisations reusing the data, and needed to be made available and easily accessible to all with minimal barriers.

Emphasis was put on the need for the data to be machine-readable and understandable, with good supporting documentation and metadata so as to aid its reuse. Reference was also made to the 2015 Local Government Transparency Code and that public-sector bodies had statutory obligations to make certain data available.

The mantra that data should be "as open as possible, as closed as necessary" highlights that although a lot of data may be sensitive, using appropriate privacy enhancing techniques such as anonymisation could enable that data to be released and reused. A focus group participant added:

> *"Importantly, I think it's a part of open data that it should be accompanied by documentation or guides, like a key or a legend of some kind, because a meaningless stream of numbers isn't really opened up. You have to know what it is, what it's related to, how it's collected, how it's divided up or labeled."*

It was highlighted that although the emphasis tended to be on public-sector organisations publishing open data, private-sector organisations should be encouraged to publish open data too, however this may prove difficult.

Current open-data use

Below are different data publishers, data types and datasets currently being used by the organisations and individuals that took park in the consultation, grouped by ODM into overarching themes. These responses came from both the survey participants, as well as focus-group attendees.

Socio-economic

- Housing affordability and home ownership
- Community assets
- Office of National Statistics census data, including unemployment rates, demographics, ethnicity, religion, population size
- Index of Multiple Deprivation
- Referral rates to community services
- Benefit claims

Environmental

- Waste licences, permits and exemptions
- Weather data
- Green space
- Flood data

Health

- COVID-19
- GP data and prescription rates
- Overseas health data
- Emergency hospital admissions
- Neighbourhood-level data on obesity and mental health

Mobility

- MOT data
- TfL data
- STATS19 road-traffic-collision data
- Public transport links

Geo-spatial

- Land Registry
- OpenStreetMap
- Ordnance Survey
- Postcode lookups
- Light Detection and Ranging (LIDAR) topographical data
- Green belt
- Strategic housing and new-development sites

Education

- OFSTED results
- Measures of childhood development
- School attendance data
- Early Years ASQ scores
- Key Stage 2 and 4 data

Safety and security

- Police crime data
- Crime Survey data
- Ambulance data

Regulatory

- Companies House data
- Licensed premises data
- US administrative data

Democracy

- Local authority spending data

Crowdsourced data

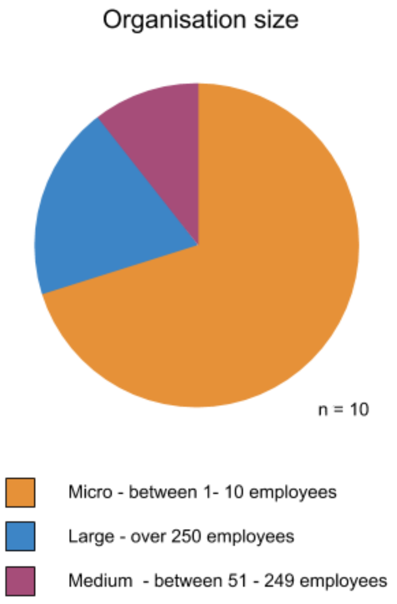
- Consultation platform data

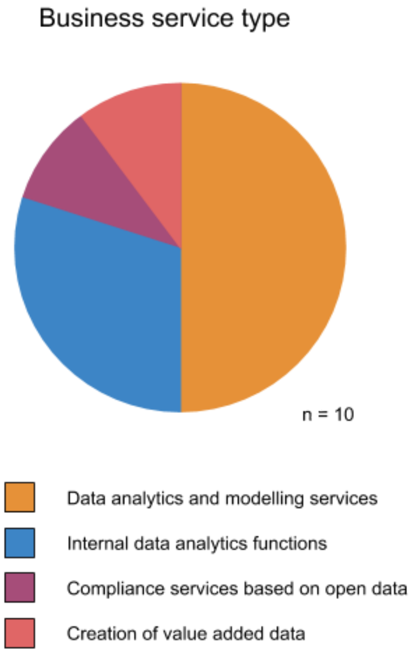
Other identified data – not necessarily open data

- Cyber Security Breaches Survey
- Primary care data anonymised with permissions in place
- Hospital records anonymised with permissions in place
- Prison healthcare records with permissions in place
- Clinical Practice Research Datalink
- National Cancer Registration and Analysis Service
- Hospital Episode Statistics
- Twitter and Instagram
- Health and leisure facilities data - although organisations were being encouraged to publish this

Applications of data – business sector

Of the business cohort, seven were based in or proximate to GM, one based in Liverpool, one in the Isle of Man and one in Milton Keynes. The cohort was predominantly micro-businesses (7) with large (2) and medium-sized organisations (1) represented.





Services within the cohort group using open data can be broken down into four broad categories:

- Organisations offering data analysis services (5)

- Organisations using data to support the aims of their businesses (3)
- Compliance services based upon open data (1)
- Creation of value-added data (1)

One organisation, which produces web-based mapping for decision support, said they use a broad selection of data: social, economic, environmental, infrastructure and transport data to understand the spatial impacts of climate change. A large part of their work is to identify and assess flood risk through modelling of public-sector estates using data from the Department for Environment, Food and Rural Affairs (DEFRA), Environment Agency, Ordnance Survey, as well as European LIDAR data. For work within Greater Manchester, the GM Infrastructure Map is used to help identify strategic housing sites and new developments for flood-risk assessment.

Another data analytics business uses open data to support utilities companies through development of financial and carbon-reduction models, as well as proof of concept work. Data from sources such as the Driver and Vehicle Licensing Agency, Land Registry and TfL is used at present as well as COVID-19 data. Similarly a planning consultancy uses Land Registry and Companies House data to identify land ownership to aid with infrastructure planning applications.

Among organisations that are using open data to support the aims of their business is the Cooperative Group. It has developed the Wellbeing Index to provide insight into the wellbeing of over 28,000 communities across the UK. By entering a postcode, the index allows users to view scores across many different measures – from the quality of education, housing affordability and public transport links, to the amount of green space and the number of community centres – providing a useful snapshot of the strengths and challenges facing that community. This insight tool is used to inform the Cooperative Group’s community strategy and identify areas of need. Since its development, the index has been made openly available and is used by external stakeholders such as local authorities, charities and social researchers.

Data is integral to the insurance industry, sources such as the Driver and Vehicle Standards Agency’s MOT Application Programmable Interface (API) allows the MOT status of vehicles to be accessed, and datasets such as STATS19 data allow identification of collision hotspots. Current research is looking at the impact of COVID-19 on vehicle mileage through MOT data. The use of topographic LIDAR information coupled with Google Images allows assessments to be made of flood risk.

The Environment Agency open ePR (electronic Public Register) API has created an opportunity for a Manchester-based business to develop compliance services that allow waste company licences, permits and exemptions to be checked to see if they are lawfully allowed to transport, recycle or dispose of different categories of waste.

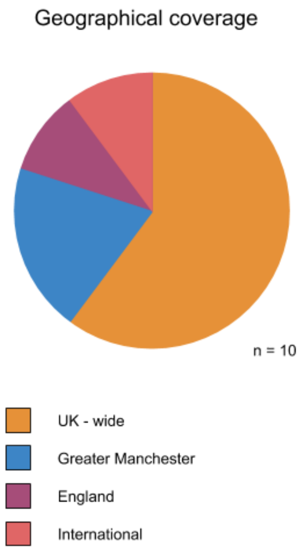
Open data creates a low-risk environment for ideas to be explored. In a brainstorming session, a group of developers working for a popular hotel booking site wanted to explore whether weather has an effect on booking behaviours. Location data was provided by Google Analytics and, although the idea never came to fruition due to lack of access to historical weather data, having access to weather data through an API allowed the concept to be explored.

Geography

Open data by its very definition is not constrained to a particular geographical territory. This enables applications and services developed in one place to be scaled or applied in others – just as long as the data structure and coverage is consistent.

A good example of this is that services developed on top of Transport for Greater Manchester's (TfGM) transit schedule have been internationally adopted in the General Transit Feed Specification (GTFS). When the feed was made available in 2012, a Berlin-based developer, Stefan Wehmeyer, almost immediately incorporated the feed into his Mapnificent application, allowing anyone to see travel times to and from locations within GM. The GTFS feed also enabled GM to become one of the first locations outside London to be integrated into the CityMapper application.

Within the business cohort, the majority of businesses saw their market being UK-wide (6), with two having a GM focus, one to England and one with an international focus.
The geographical coverage of the businesses' reach was also reflected in the types of data that they were accessing and the ways they were accessing it. Of the 10 businesses, three were accessing data through a local data portal, six through a national data portal and three through a third-party service.



The challenge of using local data was that it is often inconsistent and poorly described, meaning it takes more effort to use. For organisations providing country-wide services, datasets that can be subsetted down to a local geography are seen as more efficient.

Applications of data – academic sector

The use of open data within the academic cohort was split across methods and training, use of existing data by researchers, and the creation of new data through research itself, with participants predominantly based at The University of Manchester (10), along with UK Data Service (2), JISC (1), Manchester Metropolitan University (1) and the University of Leeds (1).

Open data featured heavily within healthcare research, with four participants stating that open data, coupled with more restricted healthcare data, allowed for better understanding of the prevalence of diseases and associated factors across a range of interventions on population health.

Research within the School of Health Sciences at The University of Manchester addresses specific health issues, such as sudden cardiac death, along with the evaluation of existing health services and the roll out of digital innovations. Prescription data and prescribing analytics has been used to understand the prescribing of antibiotics across local GP surgeries in Greater Manchester. Research into sudden cardiac death requires access to electrocardiogram (ECG) data so that algorithms can be developed to detect those at risk, and thus enabling timely intervention and treatment. Due to the difficulty in obtaining this data in the UK, it is acquired either from the US or China.

Using open data to develop research methods and to teach data skills is especially useful when looking at data that is relevant to GM. As students tend to be based in Manchester, being able to teach data and research techniques, with the added context that comes with living in the area, allows for better understanding and the utility of different analytical approaches. For example, by looking at licensing data from Manchester City Council and reported crimes from Greater Manchester Police, students are able to make a comparative analysis of violent crime in Manchester city centre with other locations.

Developing advanced algorithmic processes to identify hidden patterns requires access to good quality training data. Sources such as the UC Irvine Repository enable the development and training of machine-learning algorithms and other advanced analytical techniques.

Methods training can integrate other sources of data, such as crowdsourced data made available through Fix My Street or What Do They Know, data acquired through web scraping, or social media feeds like Twitter and Instagram.

Applications of data – VCSE and other

The use of open data within the VCSE sector enables organisations to identify need and target resources, as well as to create the evidence-base for programme development and funding.

The VCSE cohort comprised a data analyst working for a housing association, a trustee of an international education charity, a data-impact and evaluation advisor for a children's charity, a data and insight manager for a sports-employment charity, and a cycling campaigner. Grouped in this category was, also, an individual using local-authority spending data released under the Transparency Code, a digital lead for Manchester Active and an enterprise architect from a public broadcaster.

Identifying need, and understanding the profile of areas in which they work, was key to the participating charities, and as these are national organisations with a presence in GM, all the data they use is accessed through national data portals. The children's charity uses a broad array of data to understand the communities it is working with. Data is accessed from multiple sources, including the Office of National Statistics, Department of Education, Police.uk, and the Ministry of Housing, Communities and Local Government.

The sports-employment charity focuses on using the Index of Multiple Deprivation (IMD) to develop impact measurements and understand how it can be at the forefront of what this sector can do with the right tools. It uses datasets to find and target areas of deprivation "rather than sticking a pin in a map, going to that area and asking, is this deprived or not?". One of the KPIs developed by the charity is that two per cent of its participants must come from the two most deprived deciles of the IMD.

For the cycling campaigner, data created from recent consultations regarding the A56 cycleway in Trafford, using the Common Place platform, has enabled cycling to be pushed up the agenda through the creation of a database of categorised comments highlighting what does and doesn't work.

A variety of open data is used by the housing association so that it can understand and identify problems, and improve housing provision and local communities. Topographical data is used to identify flood risk to properties, while Energy Performance Certificate data allows it to understand the energy efficiency of the housing stock and identify potential improvements. Geographical information regarding locations of parks and cycle paths helps to promote a more active lifestyle.

Manchester Active is working closely with the Open Data Institute regarding open data for the sports and leisure industry, and to promote the use of the OpenActive data standard for the sharing of relevant opportunities. This involves working with many sports and leisure operators in Manchester and Greater Manchester so that they can make their facilities, activities, timetable and booking data openly available in a standardised way. This allows residents and organisations to find and take opportunities more easily. The rationale is that "it's easier for a Manchester resident to find their local Chinese and order their favourite dish than it is to book a squash court", and so booking sports and leisure opportunities should be as easy as using Uber or Trivago.

The Local Government Transparency Code compels local authorities to make their spending data above £500 publicly available. The individual, in their spare time, downloads this data, cleans it up, cross references the organisations involved with company and charity registries, and also adds Standard Industrial Classification (SIC) codes where appropriate to identify patterns in spending across local government in the North West.

Data that should be made available

Below are different data publishers, data types and datasets currently being used by the organisations and individuals that took park in the consultation, grouped by ODM into overarching themes. These responses came from both the survey participants, as well as focus group attendees.

There was a general consensus that the more data that is made available, the better, but across the three cohorts, specific datasets were identified by participants that would bring value to their particular organisation or practice.

Socio-economic

- More transparency on gender-pay gaps, hiring statistics, tendering and subcontracts
- Creation or release of more community data that describes the distinctiveness of Greater Manchester's communities
- Various indicators of the social determinants of health, to understand how these factors interplay and impact on the population's health – at the moment, while useful, IMD only covers some aspects
- Access to more granular and consistent data across the four nations of the UK would enable better local insight and would particularly help the development of the Cooperative's Wellbeing Index
- Access to more data around education, employment and health measures, such as attainment, school-readiness, mental-health-service admissions and unemployment by age group

Environmental

- Air-quality data is generally made openly available where it is collected, but more consistent collection standards are required
- Digitised waste-management data, including details of waste permits, which could enable identification of sites permitted to deal with relevant waste and user search by waste type, ensuring waste ends up at appropriate facilities, streamlining waste supply chains and seeing greater compliance
- Municipal waste data, which would allow proper interrogation and analysis of this at a national level, helping to spot trends and patterns
- National data on bin collections – data like this would enable innovations like the Leeds Bins App – getting people to put their bins out correctly is useful for efficiency and for preventing the need to engage private firms that may not be appropriately licensed
- Release of a national, open dataset of household-waste recycling centres

Health

- Summary statistics on GP practices
- Making available ECG data in a way that would allow the development of more effective heart treatments, especially as it is now suspected that COVID-19 could affect cardiac function and structure
- Public Health England data (e.g. alcohol and substance abuse treatment, smoking cessation services) to be linked with primary-care records
- Hospital (inpatient, outpatient) administrative data including diagnosis, treatments and prescriptions
- Although there is an awareness of the strict governance around health records, making them available in an anonymised form would be useful to compare new findings with routine data from typical NHS patients

Mobility

- Better public transport data in GM
- More data about cycling, granular data about cycling accidents and data feeds from cycle counters located on cycle routes, such as Oxford Road

Geo-spatial

- Access to the Postcode Address File
- Spatial planning policies, available in machine-readable format, would allow better analysis of policy decisions and to understand future land usage

Infrastructure

- Digital economy infrastructure and industrial manufacturing
- Education
- Independent schools releasing the same data as those under the local authority
- Data on Early Years Foundation Stage
- Data on Ages and Stages Questionnaire (ASQ3)
- Information relating to destinations of students post-18
- Summary statistics on the performance of schools

Safety and security

- Up-to-date police crime data for GM
- Release of emergency services and ambulance data, in line with London and Birmingham

- Alignment with the types of crime data released by the Metropolitan Police, such as the Public Attitude Survey
- Attitudes towards police perceptions of safety, similar to that covered by the Crime Survey for England and Wales
- Reports of cybercrime and fraud in GM

Democracy

- Local authority polls and surveys
- Manchester Residents Survey

Other comments

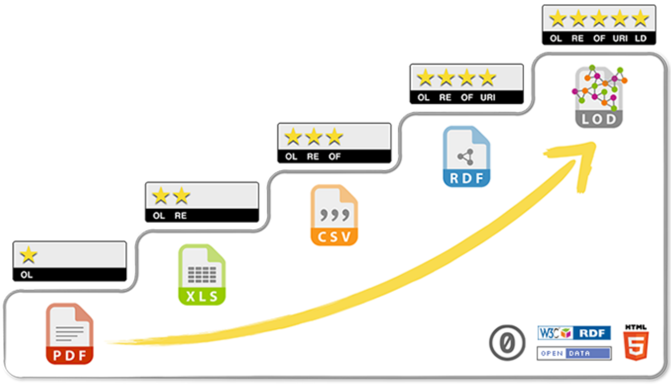
- More timely release of data, particularly when waiting for data to be released by central government
- The creation of automatically generated synthetic datasets, which would enable the demonstration of what's in the data without risk of disclosure or abuse
- Joined-up data provision across GM local authorities, in a similar way to Cheshire East Council, and Cheshire West and Chester Council

Challenges and concerns when accessing public sector data

As evidenced earlier in this report, there is a lot of activity taking place in GM that is powered by open data, but there are still concerns about the quality and stability of datasets that are being made available.

Technical

For data to be made as useful as possible, it needs to be made available in a format that enables greatest reuse. Non-proprietary formats that don't require particular software should be encouraged. Sir Tim Berners-Lee, founder of the ODI, proposed that open data should be rated on a five-star scale, with one star denoting non-machine-readable formats, and five stars where data is made available as linked open data. Comma separated values (CSV) is a non-proprietary format.



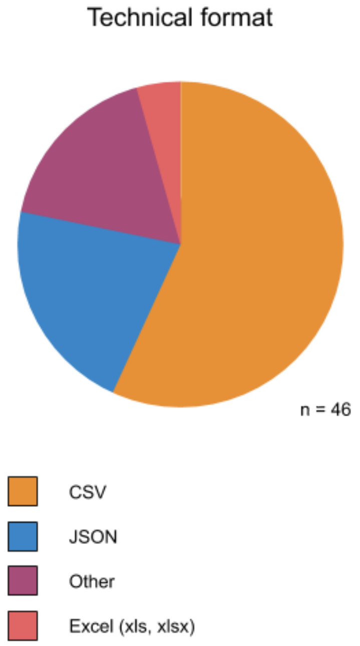
Five stars of open data - Tim Berners-Lee

Across sectors the most favoured formats for open data were:

- CSV (26)
- JavaScript Object Notation (JSON) (10)
- Other (8)

'Other' contained a variety of data formats, including:

- DICOM – Digital Imaging and Communications in Medicine
- XML – eXtensible Markup Language
- NetCDF – Network Common Data Form
- RMS – Record Management System
- DAT – Data File
- TXT – Plaintext



- WMS/WFS Web Map Service/Web Feature Service
- SQL – Structured Query Language
- Excel – Microsoft Excel

As well as making data available as a static download, it was also suggested that for certain types of data, access through an API would be preferable.

As well as using a non-proprietary format, data publishers should try to adopt and maintain commonly used open standards for data structure, such as GTFS for public-transport schedules. An example of this from the consultation is Manchester Active, which is driving the adoption of OpenActive as a common standard for the sharing of open data for sport and leisure activities in GM.

Additional data resources

In order to make the data more usable, participants stated that data needs to be published with a standard set of metadata, and supporting documentation, to enable greater understanding and increased usage. This should be backed up by a commitment to maintain the documentation and metadata through the lifecycle of the data.

Metadata requirements

Relevance	context, coverage, original purpose, granularity, summary, time frame
Usability	labelling, documentation, licence, access, machine readability, language used, format, schema, ability to share
Quality	collection methods, provenance, consistency of formatting and labeling, completeness, what has been excluded

Documentation should also be provided. The level of documentation detail will vary depending on the complexity of the dataset and the area it covers. It would typically offer:

- A high-level summary – the purpose of creation and what it describes
- Access information – how to access the data, location of archives and mirrors, if available
- Indicators – summary statistics providing insight into size, rate of growth, quality and update frequency
- Relationships – locations of other datasets that were used to construct the dataset
- Scope and coverage – description of the of the contents of the dataset, the types of entity it describes, geographic focus and the time period to which it applies
- Provenance – how the data was processed and collected prior to publication

- Technical documentation – data formats and schemas, with sample records potentially included for illustrative purposes

If data is being made available through an API, documentation that describes how to use the API will also be created.

Additionally, participants suggested that persistent URLs should be used so that links to the datasets don't break when they are updated.

Being able to preview data was considered important as it allows users to assess the data quickly so that time isn't wasted downloading data only to find it isn't suitable.

Having an up-to-date data dictionary or catalogue means that data users can easily find the data they need.

To aid with data use, it was suggested that a common directory of tools developed using the data could be shared. Further, it was proposed that there could be a way to share feedback or tips regarding the usage of the data.

Keeping data up to date and maintained is essential for any data portal or repository, but having access to historical data, so that trends can be analysed, models developed and users can get a better understanding of change over time, can enable greater insight.

Finally, it was suggested that a full-time data manager with a remit to release GM NHS data could give GM's healthcare an international advantage. For example, in privatised systems such as the United States, data is fragmented between hospitals and almost impossible to access on a city-wide or state-wide basis.

Quality

Data quality can be a big issue where data isn't as described, is missing crucial elements or doesn't conform to a particular standard. Continuity and maintenance of data is crucial to creating confidence and reducing the perceived risk of open data use, especially if an organisation is investing in the development of products and services based upon that data.

Consistency

A common problem with trying to pull together data from different suppliers is that different data describes the same things differently. It is a well-worn adage that most data scientists and analysts spend the greater proportion of their time trying to understand and clean data, prior to its actual use. Alluding to inefficiencies within the data-science profession, one respondent commented:

“I'd say that 90 per cent of the work I've done as a data scientist has been what we call ‘data munging’, where it's just moving things into a format that we can actually use. And it's very upsetting to all the people that trained to go into data science... I remember spending two weeks with a colleague just writing SQL trying to get things into a format that we needed. And the whole time I was just like, ‘why am I doing this?’ I didn't want, I don't want, a job where I'm just writing SQL for two weeks.”

The issue of data consistency was highlighted. At present there is little consistency across the 10 boroughs of Greater Manchester. More coherent and consistent planning of data would enable the development of better services and insight across the entire city-region. This would be especially relevant when undertaking strategic housing and land-availability assessments.

Continuity

As data is made available and used it creates dependencies. The data analyst working for the housing association recounted how the role of their organisation is to create livable, safe communities and access to police crime data enabled them to target interventions from home-security adaptations through to allocating resources. For the last three years, this data has stopped being published, creating challenges for the organisation.

Risk

Lack of access to good-quality open data stymies innovation, hinders experimentation and ultimately creates distrust of the data that is being released. A focus group participant stated:

“You take a problem that seems like it should be a quick win. And suddenly, it turns into, like, ‘let's spend two weeks trying to convert postcodes or whatever else’. And it's just, it's these little things, I guess, we, you know, we make assumptions that the data quality, and the data sets, are going to be good and well documented. And then

> *when they aren't, it's kind of defeating. And then we reach a point where it's like, 'let's not even tell our boss about this project', because we can't guarantee anything."*

Another respondent outlined their company's approach to developing services using open data:

> *"So we always caveat it all in the fact that it's based on open data, and the risk associated with that, and refer them back to the originator of the data early… We have to make clients aware of the fact that that data, at any time, could disappear."*

Skills

Lack of skills and time to find data that was usable was identified as a barrier to using open data effectively for a small charity. This was also highlighted in one of the larger VCSE organisations, where there was a willingness to use more data, but the capacity to do so was missing.

Health data

Accessing health data throws up a number of understandable challenges, especially as much of the data has protected status. Yet there is perceived to be:

- A lack of clear information governance (unclear procedures to obtain approvals to get access data)
- A data hoarding mentality by organisations
- An absence of secure platforms to analyse data (a lot of this data cannot be freely shared on the internet)
- More challenges around data protection the more informative the data is, such as patient-level data.

Locally collected, nationally collated data

Lack of timely access to data, especially regarding census and other demographic data, was highlighted. This was also identified as an opportunity in regards to STATS19 data, which notes road traffic collisions. Local authorities or transport agencies compile the data and submit it to the Department for Transport, but having earlier sight of the data could enable organisations to act earlier.

Overcoming challenges and recommendations

One considerable challenge identified and captured in the previous section through direct responses from the participants was the perceived or actual risks associated with using local open data.

We therefore offer a number of foundational recommendations, many of which are also technical concerns, along with more specific ones speaking to the various challenges, that we believe will start to reduce these risks.

One additional section, promoting open data use, has been added to reflect the respondents' belief that work also needs to be done to help build confidence in the open data supply.

Foundational recommendations

- Open-data infrastructure should be treated as essential infrastructure
- Develop an open-data strategy that is joined up with GM's other strategies
- Design a central repository with good user experience as the first place for GM's open data and ensure it is well maintained
- When publishing open data, create clear documentation and ensure that supporting materials and metadata is included, and kept up to date
- Maintain an up-to-date data catalogue
- Ensure that open licences relevant to the data are explicit
- Create persistent links for datasets to minimise broken data links

Technical

- Data should be made available to download in non-proprietary formats, as well as through an API if appropriate
- Ensure that a data-publication schedule is included with the data
- Allow data reusers to sample or preview the data before committing to download entire datasets
- Create a shared repository of tools built on published data

Quality

- Create a process so that organisations can become open-data compliant and release high-quality open data

Consistency

- Promote the adoption of consistent standards across GM's open data

- Create a common agreement across public-sector agencies to maintain consistency in data governance

Skills

- Curate collections of datasets of benefit to VCSE organisations that don’t have the capacity to search through different datasets

Health data

- Invest in full-time data managers with a remit to release GM NHS data
- Explore the viability of creating synthetic data for sensitive and protected data
- Develop best-practice guidelines regarding data anonymisation and other privacy-enhancing techniques
- Create a health data ‘sandbox’ to allow approved users to analyse data in a secure platform while preserving patient privacy

Locally collected, nationally collated

- Enable GM data that is sent to central government for collation and release to be made available in advance, enabling GM organisations to understand changes and better prepare for when these collated national datasets are released (e.g. STATS19 and socio-economic data)

Promoting open data use

- Build use cases and case studies - especially relating to the leisure industry - so that others can see the benefits of releasing standardised open data
- Engage with the open-data-reuse community
- Create a means to answer questions and accommodate feedback from the data-reuse community
- Promote new data releases and build interest in public-sector data, particularly by leveraging the tech community to raise awareness of published data, and supporting the development of hack events and other profile-raising activities.

Datasets

- Given the wide range of datasets that different sectors would like to be made available, effort needs to go into understanding the current data that is available and prioritising what to do next, based on value, and with the support of those who would benefit from it - participants specifically suggested data about people’s quality of life, such as proposed disruption, building work, road maintenance etc.

Conclusion and next steps

This exercise provided insight into some of the many possible uses of public-sector data within Greater Manchester. It has also set out the important and recurring open-data challenges faced by the organisations surveyed. These challenges largely relate to issues around data quality, consistency and reliability, as well as a need for the relevant skills to access and interpret open data within some participant organisations.

The recommendations outlined in this report offer some suggestions to overcome the identified challenges. These findings will be interpreted and adopted by the GMCA in its action plan to create a better information and open-data ecosystem in Greater Manchester, which will be published in early 2021.

Beyond the specific recommendations outlined above, a key message that underpins these findings is that releasing data alone does not mean it will be used – efforts must be made to reduce barriers to use, as well as promote use. Building and maintaining dialogue between data producers and data users is key to encouraging greater data re-use. This exercise was the first step in this wider and important process of user engagement.

Annex 1. Survey Questions

Q1. What sector do you work in?

Q2. Where is your organisation based?

Q3. What is the geographical focus of your organisation?

Q4. What is the principal activity of your organisation?

Q5. What is your job role?

Q6. What is the size of your organisation?

Q7. What public sector data do you currently use?

Q8. Is this data specific to Greater Manchester?

Q9. What applications do you use this data for?

Q10. Is this data critical to your work?

Q11 How do you access this data? (Please tick all that apply)

Q11.1 Local data portal (such as a local authority website)

Q11.2 National data portal (such as gov.uk)

Q11.3 Third-party data provider

Q11.4 Other (please specify)

Q12. What public sector datasets would you like made available?

Q13. What would this enable you to do?

Q14. What technical formats (such as CSV/JSON/GTFS) would make this data most usable to you?

Q15. What would you like to see provided with the data to make it more usable?

Q16. What challenges do you or your organisation face in accessing public sector data?

Q17. What would you like to see done to help overcome these challenges?

Q18. Is there anything else that you would like to add or tell us?

Q19. We will be running some online focus groups for business, academic and voluntary and charitable sectors, where participants will be able to discuss and share their open data needs in more detail. These will take place between 24-26th November 2020.Would you be interested in attending one of these focus groups?

Q20. We might want to contact you if we have a question regarding your answers in this survey. Would you be willing for a member of Open Data Manchester to contact you?

Q21. Would you be interested in receiving a copy of the final report?

Q22. If you answered yes to any of the above, please put your name and email address below. This information will remain strictly confidential and will only be used to contact you for the purposes of this consultation. For more information on Open Data Manchester's Privacy Policy.

Q22.1 Name

Q22.2 Email address