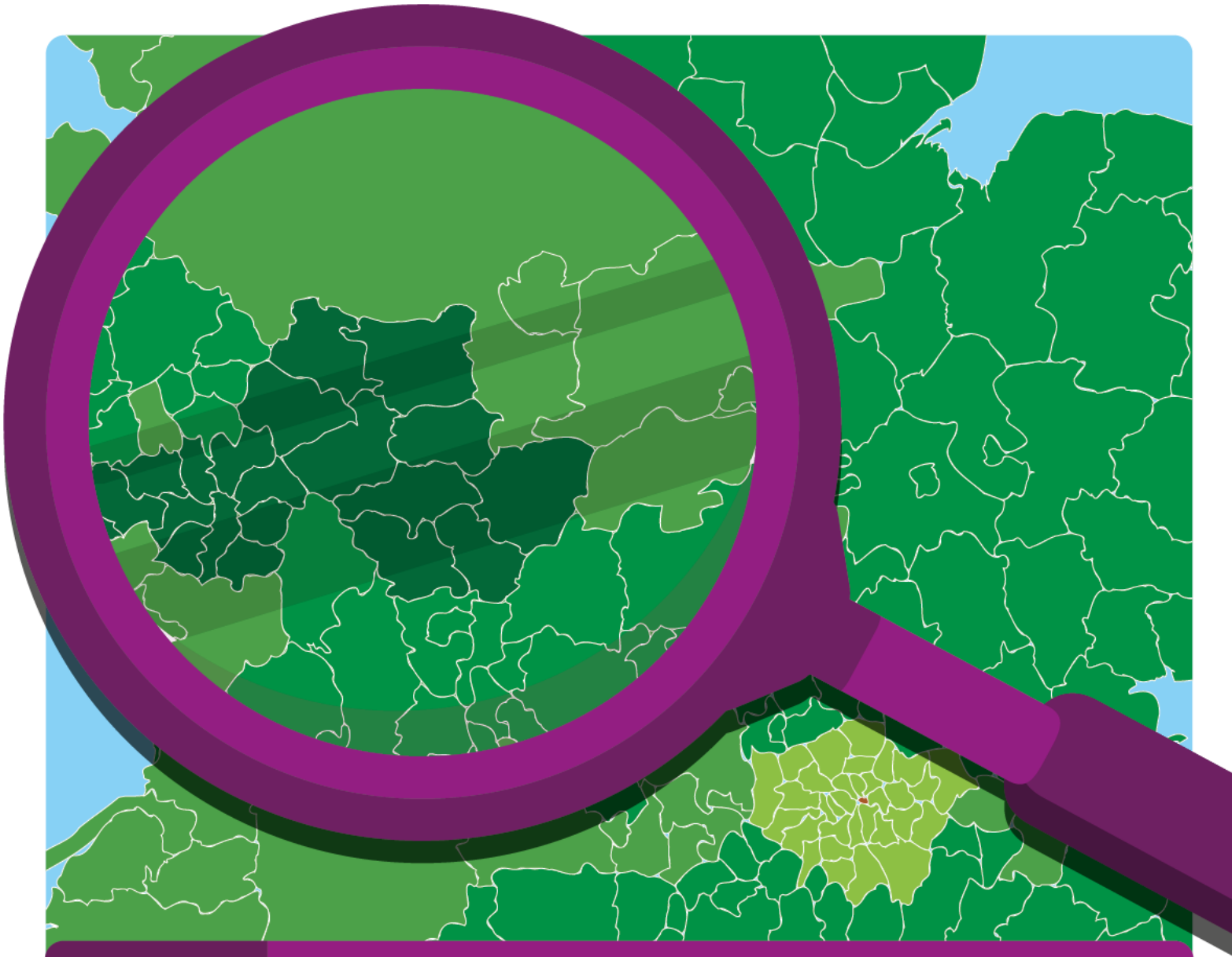
Local Government Association

Demystifying data

An overview of common data terms



Research guide

Purpose

Data terminology can sometimes be confusing to non-data professionals.

The purpose of this document is to bring commonly used data terms together into a handy single reference source so that anyone with an interest in data can better understand basic data concepts.

Further reading

For more information about data use in local government, please visit the better use of data programme pages on the LGA's website.

Data glossary

Terminology	Meaning
Artificial Intelligence (AI)	Artificial intelligence (AI) is the process of using computers to do work that has traditionally has only been able to be done by humans. This includes reasoning and solving complex problems. AI is not new and has been around since the 1950s. In 2022 Generative AI was released to the public, prompting greater interest in AI and AI-enabled capabilities. AI describes a wide range of technologies that power many of the services and goods we use every day, including chatbots and search engines.
Algorithm	A traditional algorithm is a set of rules followed in an order or sequence. They are essentially used to automate a process. Algorithms enable computer systems to work.
Application Programming Interface (API)	Application Programming Interfaces (APIs) are like a handshake that enable data to be exchanged from one system to another. Technically APIs enable communication by requesting data and sending data in response to requests. For systems to integrate with one another using APIs, there needs to be an API or connection method enabled for each system. One of the benefits

	of APIs is that they automate data transfer, removing the need for manual ways of moving data between systems.
Attribute	An attribute is a piece of data about an object or an entity. It is data that describes other data, so for example, if ‘Customer’ was the entity then gender, age and ethnicity might all be examples of attributes associated with it.
Azure	Microsoft Azure is a cloud platform made up of multiple products and services. It includes computing, analytics, storage, and networking and enables users to build, deploy, and manage applications through Microsoft-managed data centres.
Batch processing	Batch processing enables large volumes of data to be processed in batches. Instead of processing each data transaction individually, tasks are bundled up and processed together. Batch processing jobs are usually run at scheduled times and because of their size, sometimes out of hours. As an example, invoices for services might be generated by batch processing at the end of each month.
Big Data	Big Data includes very large, complex data sets that traditional tools can't efficiently manage or analyse. It covers both structured and unstructured data and can pose challenges due to its size and diversity.
Business Analyst	Business Analysts identify organisational needs and find solutions to problems. They collaborate with stakeholders such as management, IT teams, and end-users to ensure business goals align with technical capabilities. Their role includes examining data, processes, and systems to enhance efficiency and productivity. They use data in their work.
Business intelligence (BI)	Business intelligence (BI) means transforming raw data into meaningful insights that enable strategic decision-making.

ChatGPT	ChatGPT is an AI tool that works like a chatbot and can respond to questions asked of it by looking at vast quantities of text data to understand patterns, and responding with the pattern it considers most appropriate. ChatGPT is an example of a Large Language Model (LLM).
Classification	Data classification is the process of putting data into meaningful categories based on shared characteristics or attributes. It can help to transform unstructured data to structured data.
Cloud	Cloud computing refers to the delivery of various services over the internet. This includes storage, databases, servers, networking and software. Cloud services are not actually based in a ‘cloud’. They are hosted on large remote servers located in physical data centres and connected via the internet.
Clustering	Clustering or cluster analysis involves grouping data based on specific characteristics. As an example, you might have a dataset that includes ‘customer’ and against each customer a record of the services that they use. Clustering enables you to select a characteristic – like ‘prefers online services’ and cluster data together on that basis.
Correlation	Correlation describes the extent to which two variables are related.
Comma Separated Values (CSV)	A CSV (Comma-Separated Values) file stores data in a tabular format, where each line represents a data record. Each record consists of one or more fields separated by commas. For example: Name, Email, Phone Number, Address. CSV files can be used to enable data exchange between different applications.
Dashboard	A data dashboard is a visual tool that organises and displays key data for easy understanding, pulling information from a database to enable real-time monitoring, analysis, and decision-making.

Data analytics	Data analytics is the process of inspecting, cleaning, transforming, and modelling data to uncover valuable insights. It involves using various analytical techniques and tools to interpret raw data, finding patterns, trends, and relationships within a dataset.
Data architecture	Data architecture defines how organisations gather, store, process and manage data. It is a practice that includes defining the tools, policies, and standards for data handling. Data Architects design how data flows from different sources to enable processing, distribution, and visualisation. Their purpose is to ensure data meets organisational needs effectively.
Database	A database is a collection of structured data, typically stored electronically. It is usually managed by a database management system (for example SQL Server), which acts as an interface between the database and its users or programs, enabling efficient retrieval, updates, and management of the data.
Database schema	A database schema is a blueprint defining how data is organised in a database. It includes table names, fields, data types, and relationships between entities, serving as a framework for data storage and management.
Data catalogue	A data catalogue is a comprehensive list or catalogue of all data assets within an organisation. It is intended to help data professionals in efficiently finding the most suitable data for analytical purposes.
Data centre	A data centre is a physical site that includes the computing infrastructure needed for IT systems, such as servers, data storage drives, and network equipment.
Data cleansing	Data cleansing is the process of identifying and correcting, or removing, errors, inconsistencies, and inaccuracies within a

	dataset. This ensures that the data is accurate, consistent, and reliable, which is critical for effective analysis.
Data custodian	Data custodians are responsible for capturing, storing and disposing of data in line with data owner requirements and standards. Custodians help data stewards with technical data queries and provide advice to data teams and technical teams to ensure data standards are used. An example of a data custodian in local government would be the Local Land and Property Gazetteer (LLPG) Custodian, the main person with responsibility for maintaining the LLPG within national standards, submitting data to the national GeoPlace hub.
Data democratisation	Data democratisation is the process of enabling everyone, regardless of their technical ability, to work with data comfortably, engage in discussions about it confidently, and ultimately make data-informed decisions.
Data governance	Data governance focuses on managing data quality, security (including access), and availability. It ensures data integrity and security by defining policies, standards, and procedures for data collection, ownership, storage, processing, and use.
Data lake	A data lake is a central repository for storing large amounts of raw data in its original form. It can hold structured, semi-structured, or unstructured data. Data lakes are flexible and scalable, suitable for big data analytics, machine learning, and real-time analytics.
Data mesh	Data mesh approaches focus on a decentralised approach to data and on data as a product itself, rather than as the by-product of a process. Meshes enable services to manage data like they would any other software application or product. A data mesh approach can provide greater ownership of data at a

	service level, however, can be complex to introduce, embed and oversee. They are best used by organisations with an existing good level of existing data maturity.
Data migration	Data migration is the process of moving data from one system, storage, or computer processing environment to another.
Data mining	Data mining is the process of discovering patterns in large data sets using methods from machine learning, statistics, and database systems.
Data modelling	Data modelling involves creating visual representations of data structures, relationships, and rules.
Data owner	A data owner has responsibility for certain sets of data. Data owners are usually service leads (e.g. Director of Adult Social Care) and they are responsible for ensuring that the data being put into systems is accurate and correctly managed across various platforms and business processes.
Data pipeline	A data pipeline is a set of tools and processes for collecting, processing, and delivering data from one or more sources to a data platform where it can be analysed.
Data platform	Data platforms are technologies used to collect and analyse large amounts of structured and unstructured data. They store, manage, process and analyse data and can enable data to be visualised. A data platform can be a single tool or combine multiple components. When teams talk about ‘We have an Azure data warehouse’ they are talking about a specific type of data platform.
Data quality	Data quality refers to the reliability, accuracy, completeness, and consistency of data. High-quality data is devoid of errors, inconsistencies, and inaccuracies, and is suitable for decision-making and analysis.

	You can find out more about data quality in The Government Data Quality Framework - GOV.UK
Data science	Data science uses maths, statistics, computer programming and advanced analytics, AI and machine learning, along with subject matter expertise to find actionable insights in an organisation's data.
Data security	Data security safeguards digital information against unauthorised access, corruption, damage, and theft.
Data silo	Data silos are isolated collections of data, inaccessible by other groups within the same organisation.
Data standards	A data standard is an agreed and consistently defined way of describing data, or the 'rules' associated with a particular data entity. Data standards help to improve data quality and consistency. By using common formats, it is easier to match, merge and share data, and for systems to integrate. An example of a data standard is the Unique Property Reference Number (UPRN) used by councils as a standard for address information.
Data steward	Data stewards are responsible for collecting, maintaining, and protecting the data assets in an organisation. The role is usually combined with a service role – so for example a Team Leader in a contact centre may be a data steward for a particular set of data, supporting the Data Owner to ensure it is well managed and maintained.
Data warehouse	A data warehouse is a centralised repository that aggregates data from various sources. Data warehouses help organisations to conduct advanced analytics, business intelligence, and machine learning tasks effectively. They are different to data lakes in that the data within them is structured.

Data Analyst	A data analyst is responsible for collecting, cleaning, and interpreting data sets to support with solving problems and making decisions that are informed by data.
Data Protection Impact Assessment (DPIA)	A Data Protection Impact Assessment (DPIA) is a risk assessment designed to identify, analyse and mitigate security risks that come with collecting, processing, using, storing, and sharing user data. Data Protection Impact Assessments (DPIAs) | ICO
Data Protection Officer (DPO)	A Data Protection Officer ensures compliance with data protection laws including the General Data Protection Regulation (GDPR). It is a requirement that councils have a nominated DPO. DPOs must be independent, have expertise in data protection and have adequate capacity to fulfil the role. DPOs should have a direct route to senior management.
Deep learning	Deep learning uses complex neural networks to simulate the human brain, enabling computers to autonomously uncover patterns and propose insights and decisions from large amounts of data.
Entity	An entity is a real world ‘thing’ that can be distinctly identified. Examples might be a customer, an employee or a building. Entities then have attributes associated with them.
Extract, Transform and Load (ETL)	ETL stands for Extract, Transform, Load, and is a fundamental process in data warehousing. This process includes extracting data from multiple sources, transforming it into a suitable format, and loading it into a data warehouse.
Fabric	Fabric is an integrated analytics platform provided by Microsoft combining different products including Azure Data Factory, Azure Synapse Analytics, and Power BI. It is used by data professionals to bring data together and enables powerful analysis.

General Data Protection Regulation (GDPR)	Data protection laws regulate how organisations use personal information. In the UK, data protection follows the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018.
Generative AI	Generative artificial intelligence, often called generative AI or gen AI, is a type of AI that can create new content like conversations, stories, images, videos, and music. An example would be typing ‘draw a pink unicorn driving a red car’ into ChatGPT and the AI creating the relevant image.
Heatmap	A heatmap uses colour to represent data, making complex information easy to understand.
Human-in-the-loop	Human-in-the-loop involves humans being ‘in the loop’ alongside artificial intelligence. Most commonly human in the loop means that humans will assess the output from AI-derived insights to take the final decisions on something. For example, AI may recommend a risk profile and potential interventions to support a vulnerable adult, but final decisions on care and support will be made by a human.
Integration (data)	Integration combines data from different sources into a single format for analysis, operations, and decision-making.
Large Language Models (LLMs)	A large language model, or LLM, is a type of generative AI system that can process and respond to human language text and prompts.
Lineage	Data lineage is the tracking of data from its origin to its destination, including all transformations. It documents the entire journey of a specific piece of data.
Machine learning (ML)	Machine learning (ML) is a branch of AI that enables computers and machines to learn from data, perform tasks autonomously, and enhance their performance and accuracy over time through experience and exposure to more data.

Master Data Management (MDM)	Master Data Management (MDM) involves creating and maintaining a single, consistent master record for each key entity within an organisation. In councils you may hear this referred to as a ‘single view of customer’ and ‘single view of asset’ for example. The process of Master Data Management includes deduplication, reconciliation, and enrichment of data to ensure a unified view of essential business information. MDM helps the management and sharing of data across departments, applications, and systems, resulting in more accurate reporting, minimising errors, and enabling well-informed decision-making.
Mean	The mathematical mean is a common way of defining an ‘average’. It is created by summing all of the values and dividing them by the total by the number of values in the set.
Median	The median is another way of defining average. It is the middle number in a set when the numbers are ordered from smallest to largest.
Metadata	Metadata is data about data, providing information to organise, manage, and understand primary data. It is crucial in digital files, libraries, websites, and databases for data management, retrieval, and organisation. Examples of metadata might include things like ‘creation date’ and ‘author’ or ‘last modification date’ as well as access permissions. You are using ‘permissions’ metadata when you give access to files in SharePoint for example.
Mode	The mode is also a way of defining average. It is the number that appears most often in a set of numbers, also called the modal value.
PowerBI	PowerBI is a Microsoft tool for visualising data from multiple sources, creating relevant dashboards.

Qualitative data	Qualitative data is non-numerical information that captures human experiences, attitudes, and behaviours. Interview transcripts or recordings are an example of qualitative data.
Quantitative data	Quantitative data is numerical information that can be measured, counted, and statistically analysed. It helps quantify variables, find patterns, and predict outcomes. An example would be the number of bins delivered per day.
Register of Processing Activities (RoPA)	Article 30 of the GDPR requires controllers and processors of personal data to keep a Record of Processing Activities (RoPA). This document outlines all types of personal data controlled and processed by an organisation. The LGA's LGInform Plus tool includes a cost-effective template for councils to use – find out more at https://vimeo.com/1046384076 and https://lginformplus.org/.
Reproducible Analytical Pipelines (RAP)	Reproducibility is important as it ensures that data analysts get the same results as each other when using the same data and methods. Reproducible Analytical Pipelines (RAP) use software engineering practices to improve reproducibility and to reduce the time analysis can take.
Sample	Sample size refers to the number of observations or individuals in a study. A suitable sample size is crucial for accurate, reliable research findings.
Senior Information Risk Owner (SIRO)	A SIRO is responsible for ensuring risk management processes and systems are in place to enable secure and effective information governance practice. The SIRO role is mandatory for public sector organisations.
Statistical significance	Statistical significance helps figure out if results are genuine or happened by chance. It's crucial in hypothesis testing and drawing reliable conclusions from data.

Structured Query Language (SQL)	SQL is a programming language for managing data in a relational database, which stores information in tables with rows and columns. Developers use SQL statements to store, update, delete, search, and retrieve data. Data analysts and developers learn and use SQL because it integrates well with different programming languages.
Structured data	Structured data is organised in a database with a fixed schema (a design that defines how data is organised), fitting into rows and columns like names and phone numbers. Unstructured data does not have this fixed scheme.
Tableau	Tableau is software for exploring, managing, and sharing data insights from different sources and enables users to create relevant dashboards.
Unstructured data	Unstructured data is information without a predefined model or structure, making it hard to collect, process, and analyse with traditional tools. Unlike structured data organised in formats like rows and columns in a database, unstructured data comes in various forms and formats. Emails are an example of unstructured data.
Variable	A variable is a data item that can assume different values. A data variable is any characteristic, number, or quantity that can be measured or counted. For example, the price of each item.
Web scraping	Web scraping is an automated method used to extract data from websites. This data is typically unstructured and is then converted into structured data. Some websites have terms of use that prohibit scraping, and scraping personal data or intellectual property without consent can lead to legal issues. Commercial companies scrape social media sites to measure public sentiment about their products for example.

Local Government Association

18 Smith Square

London SW1P 3HZ

Telephone 020 7664 3000

Fax 020 7664 3030

Email info@local.gov.uk

www.local.gov.uk