Department for
Science, Innovation
& Technology

Guidelines and best practices for making government datasets ready for AI

Helping develop effective data structures and data dissemination methods for AI readiness

January 2026

Contents

Executive summary

The United Kingdom is at a critical inflection point in its adoption of artificial intelligence across sectors. While advances in machine learning, Generative AI capabilities, and Agentic AI capabilities continue at pace, the effectiveness, safety, and legitimacy of AI adoption remain fundamentally constrained by the quality, structure, and governance of underlying data. As articulated in the AI Opportunities Action Plan[1], the realisation of AI driven public value depends not only on computational capability, but on the availability of accessible, consistent, high quality, and trustworthy datasets.

In His Majesty's Government (HMG), data collection has often prioritised operational delivery, reporting, or compliance, while considerations around reusability for advanced analytics or AI have sometimes received less attention. As a result, data often remains siloed in respective department systems and with inconsistent documentation. Many datasets lack sufficient metadata, making them difficult to repurpose. The National Audit Office has highlighted[2] that providing raw data or basic APIs without information on data quality or provenance can lead to misunderstanding and misuse, especially in AI capabilities that learn at scale without context. These risks are significantly amplified in AI capabilities, which learning patterns at scale without contextual awareness.

This document provides the initial guidelines for releasing government datasets for AI applications and sets out **guidelines, success criteria**, and **best practices** for UK public sector organisations, providing clear steps for preparing datasets to support various AI capabilities and promoting responsible data stewardship. Implementation pillars and actionable principles are included to ensure responsible AI adoption with legality, security, and public confidence.

These guidelines are intended to support:

- **Data and AI practitioners** – ensuring datasets are technically fit for AI by structuring, documenting, and governing them effectively, including developing transparent, reusable data structures and dissemination methods that enable safe and effective AI development.
- **Policy leaders and decision makers** – ensuring data use is lawful, ethical, accountable, and aligned with public trust.

[1] UK Government, *AI opportunities action plan*, published 2024, accessed 15 January 2026, https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

[2] National Audit Office, *Improving government data: A guide for senior leaders*, published July 2022, accessed 15 January 2026, https://www.nao.org.uk/wp-content/uploads/2022/07/Improving-government-data-a-guide-for-senior-leaders.pdf

As the AI landscape evolves, these guidelines will be updated to reflect emerging challenges and best practices. We welcome feedback to inform future revisions. Please share comments, suggestions, or case studies via dataAIReadiness_incubator@dsit.gov.uk.

Introduction

The value of any AI capability generally depends on five key questions:

- Does it have the right data?
- Does it perform well?
- Does it meet the business need?
- Is it cost effective?
- Is it easy to maintain and compliant?

This document looks to explore the characteristics of whether available data is suitable for AI capabilities – i.e. **does it have right data?** Effective data readiness requires consideration of the specific AI capability in use; whether traditional machine learning methods, large language models, or emerging trends in AI[3]. The quality and suitability of data depend not only on the technology and its intended use, but also on how data is defined, collected, and captured at source. Different AI use cases require tailored data quality metrics and governance arrangements to assess whether data is fit for purpose, to track data quality over time, and to ensure that appropriate controls are in place so AI capabilities can operate effectively and responsibly. Common AI use cases include:

- Generating content: e.g. producing reports, code, images, or videos
- Understanding & summarising information: e.g. documents, emails or logs
- Decision making & recommendations: e.g. identifying next best actions or providing analytical insights
- Conversational assistance: e.g. supporting users through copilots and intelligent agents to accelerate work
- Predicting outcomes: e.g. forecasting risk or future events
- Improving and governing data itself: e.g. identifying data quality issues, classifying and managing digital records, supporting retention and deletion decisions, and anonymising, aggregating, or labelling data so that users with different access permissions can safely work from a shared, trusted source of data.

Data is a foundational property that determines whether an AI system performs well and aligns with business needs. An AI-ready dataset is not defined solely by technical format but by its **context, governance, interoperability**, and **suitability** for specific AI use cases. Datasets should be self-describing or contextually complete and managed as strategic data products requiring ongoing oversight.

[3] UK Government Digital Service, *Artificial Intelligence Playbook for the UK Government*, published 10 February 2025, accessed 15 January 2026, https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government/artificial-intelligence-playbook-for-the-uk-government-html

As set out in the Open Data Institute (ODI) A framework for AI-ready data[4], a dataset may be considered AI-ready when it addresses these four components:

- **Technical optimisation**: data is structured and formatted for efficient use by machine learning systems and interoperability with AI tools.
- **Overall quality and adherence to standards:** data is accurate, complete, consistent, and maintained through effective processes that cover structure, metadata, quality monitoring (live) and dissemination
- **Legal and regulatory compliance:** data is fully aligned with applicable laws and regulations
- **Responsibly managed:** data is managed ethically and securely throughout its lifecycle

To identify current challenges and good practice, a number of interviews were conducted with public sector bodies, departments and expert organisations to surface real-world barriers, practical enablers, and common patterns, and to use these insights to shape the framework, principles, and action plan set out in this guidance.

The **AI-ready data action plan** section contains suggestions and ideas for developing AI-ready datasets and **Appendix C** contains broader details of productive insights.

[4] Open Data Institute, *A framework for AI-ready data*, published May 2025, accessed 15 January 2026, https://theodi.cdn.ngo/media/documents/A_framework_for_AI-ready_data.pdf

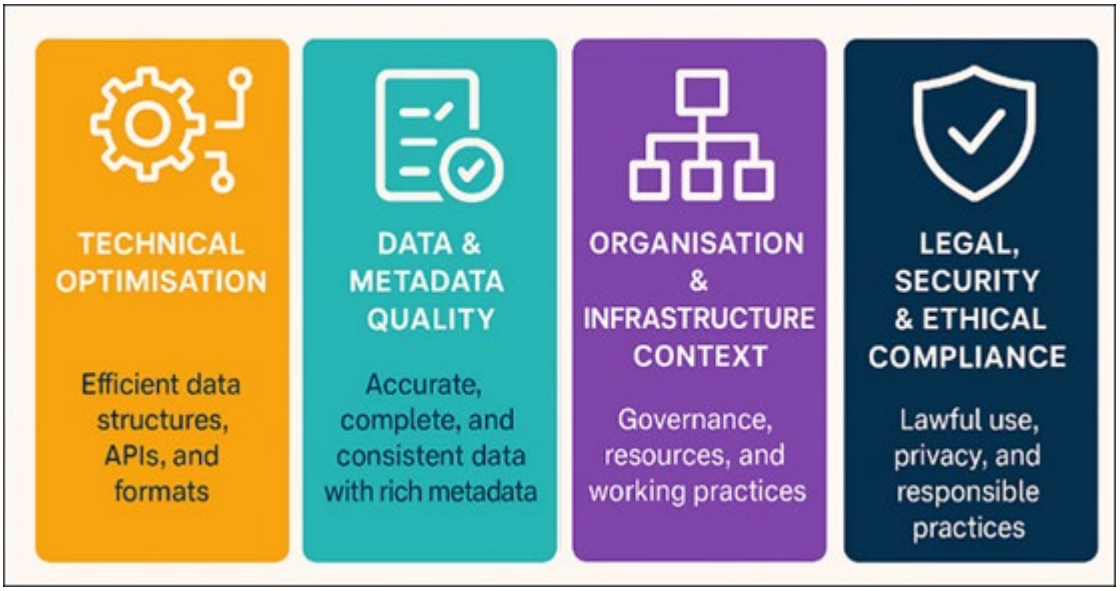
Foundations of AI-ready datasets

For data to be AI-ready, organisations will need to ensure their datasets adhere to a set of guiding principles to support its use and dissemination with confidence.

These principles support a holistic approach that balances technical rigor, ethical responsibility, and organisational readiness. They are grouped into four pillars, forming the basis for reliable AI solutions and which are based on the ODI's framework for AI-ready data:

The four pillars:

- Technical optimisation
- Data & metadata quality
- Organisation & infrastructure context
- Legal, security & ethical compliance



The four pillars of AI-ready data

The mapping between the foundational pillars and the AI-ready data principles is given in the table below:

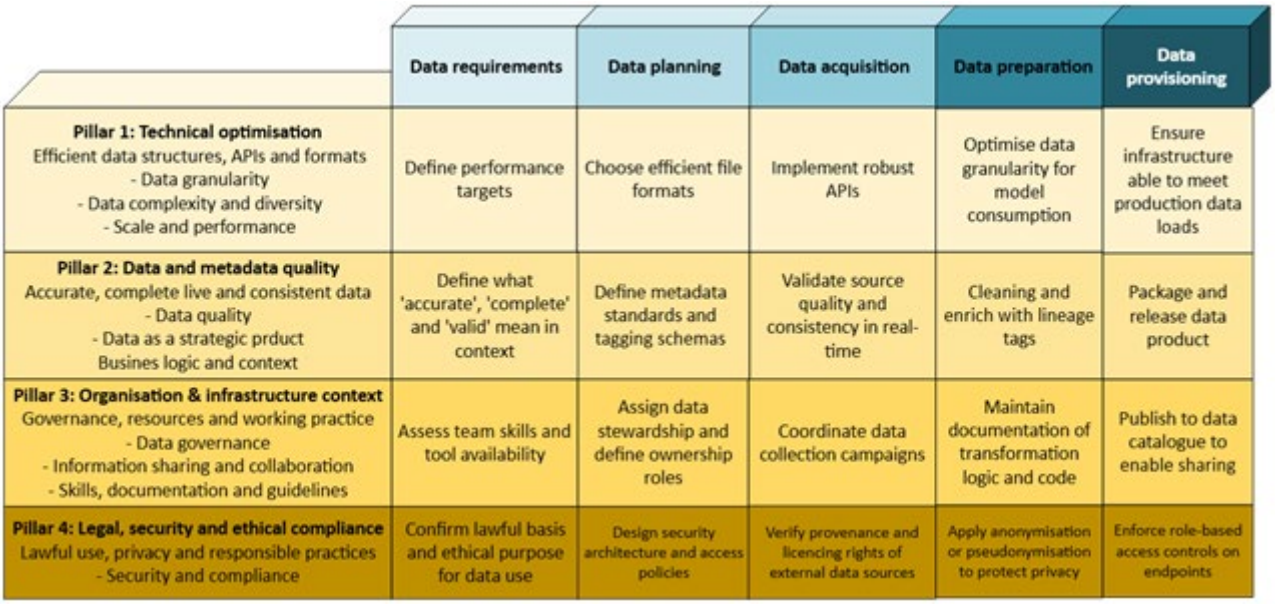
Foundational pillar	**AI-ready data principles**
Pillar 1: Technical optimisation Efficient data structures, APIs and formats	Data granularity
	Data complexity and diversity
	Scale and performance

Foundational pillar	AI-ready data principles
Pillar 2: Data and metadata Quality Accurate, complete, live and consistent data with rich meta data	Data quality
	Data as strategic product
	Business logic and context

Foundational pillar	AI-ready data principles
Pillar 3: Organisation & infrastructure context Governance, resources and working practice	Data governance
	Information sharing and collaboration
	Skills, documentation and guidelines

Foundational pillar	AI-ready data principles
Pillar 4: Legal, security & ethical compliance Lawful use, privacy and responsible practices	Security and compliance

These four pillars form the foundation for defining and managing AI-ready data. The framework applies across the entire data lifecycle and as shown the accompanying diagram, aligns with the stages set out in the ISO/IEC 8183 AI - Data lifecycle framework. It also provides example actions that correspond to each pillar.



	Data requirements	Data planning	Data acquisition	Data preparation	Data provisioning
Pillar 1: Technical optimisation Efficient data structures, APIs and formats - Data granularity - Data complexity and diversity - Scale and performance	Define performance targets	Choose efficient file formats	Implement robust APIs	Optimise data granularity for model consumption	Ensure infrastructure able to meet production data loads
Pillar 2: Data and metadata quality Accurate, complete live and consistent data - Data quality - Data as a strategic prduct Busines logic and context	Define what 'accurate', 'complete' and 'valid' mean in context	Define metadata standards and tagging schemas	Validate source quality and consistency in real-time	Cleaning and enrich with lineage tags	Package and release data product
Pillar 3: Organisation & infrastructure context Governance, resources and working practice - Data governance - Information sharing and collaboration - Skills, documentation and guidelines	Assess team skills and tool availability	Assign data stewardship and define ownership roles	Coordinate data collection campaigns	Maintain documentation of transformation logic and code	Publish to data catalogue to enable sharing
Pillar 4: Legal, security and ethical compliance Lawful use, privacy and responsible practices - Security and compliance	Confirm lawful basis and ethical purpose for data use	Design security architecture and access policies	Verify provenance and licencing rights of external data sources	Apply anonymisation or pseudonymisation to protect privacy	Enforce role-based access controls on endpoints

Pillar 1: Technical optimisation

This pillar focuses on making datasets and data services efficient, scalable, and easy to integrate into AI capabilities. This pillar addresses principles including the practical requirements of scale, interoperability, performance, and integration, recognising that AI capabilities place significantly different demands on data infrastructure compared to traditional reporting or statistical use.

The three principles which fall under this pillar are:

Data granularity

Data granularity should be managed at the level appropriate to its AI capability:

Different AI capabilities can require data at various levels of granularity, with a single representation rarely being sufficient to support all use cases. For example, fraud detection uses detailed logs, while reporting relies on aggregates. Strictly separating granular and aggregated data can cause business logic to drift and create data debt. AI-ready datasets should therefore be maintained at multiple levels of granularity (called grains), with each representation governed and documented explicitly.

The use of synthetic data is also important. Depending on the stage of an AI project such as initial development, testing, or where access to real data is restricted synthetic data may supplement or even exclusively replace real data to accelerate readiness and ensure privacy.

Essential capabilities for AI readiness:

- Maintain datasets at multiple grains mapped to model families (e.g., timeseries granular logs for ML; aggregated partitions for reporting). This should be done before getting into the feature selection process.

- Adopt a Lakehouse[5] pattern with Bronze (raw), Silver (cleaned/enriched), and Gold (aggregated/curated) layers. This approach supports the requirements of ACID[6] (Atomicity, Consistency, Isolation, and Durability - principles that ensure data remains accurate, reliable, and protected from errors during processing) to provide robust transactions as data progresses through successive stages of validation and transformation, ultimately being stored in a format optimised for high performance analytics and AI capabilities. Maintain flexibility so data can be recomposed as needs evolve. Gold layers are optimised for analytics and dashboards, but generally unsuitable for AI capabilities training, as aggregation removes the detail and variance needed for learning. For AI, access to detailed bronze or silver data is essential for training. For inference (making predictions in production), Gold data may be sufficient or preferable due to its efficiency for reporting and access.
- Use a Feature Store to dynamically generate both granular and aggregated views as needed for different AI capabilities (such as prediction, generation, understanding, decision/recommendation, and conversational/agentic tasks). This ensures all data views share the same mathematical lineage and remain consistent.
- Ensure robust data management principles atomicity, consistency, isolation, and durability are applied as data moves through validation and transformation.
- Provide explicit documentation that maps each data grain (level of detail or aggregation) to its appropriate AI use cases, acknowledging that no single dataset structure can satisfy all requirements.

Example: HMRC's approach to council tax[7] data illustrates implementing some of the data granularity principles in practice. The dataset is available in both detailed and summary format and it uses open format for data hosting structured data in the form of CSV and Parquet with greater consistency and quality, enabling its reuse across a range of analytical and AI capabilities while maintaining appropriate safeguards.

Data complexity and diversity

Public sector data environments are increasingly complex, encompassing structured records, semi structured logs, unstructured text, images, voice, biometrics and geospatial data. AI-ready datasets must explicitly support this diversity and define clear conversion, validation, and integration pathways.

Essential capabilities for AI readiness:

- **Accommodate diverse data types**: AI-ready datasets must support a wide range of data types, including structured records, unstructured text, images, geospatial data, and

[5] Microsoft, *What is the medallion lakehouse architecture?*, last updated 4 September 2025, accessed 15 January 2026, https://learn.microsoft.com/en-us/azure/databricks/lakehouse/medallion
[6] Microsoft, *ACID Properties*, last updated 6 January 2021, accessed 15 January 2026, https://learn.microsoft.com/en-us/windows/win32/cossdk/acid-properties
[7] Sam Trendall, *HMRC algorithm for automated council tax valuations offers 'greater consistency and quality'*, *PublicTechnology*, published 6 August 2025, accessed 15 January 2026, https://www.publictechnology.net/2025/08/06/economics-and-finance/hmrc-algorithm-for-automated-council-tax-valuations-offers-greater-consistency-and-quality/

critically vector embeddings. For advanced AI capabilities such as semantic search or Retrieval Augmented Generation (RAG), it is not enough to simply store raw text or images. These assets should be chunked and converted into vector embeddings[8], which are then stored in a vector database (i.e. breaking data into parts and representing each as a set of numbers that capture its meaning, enabling efficient search and retrieval for AI capabilities). This enables efficient similarity search and retrieval, making legislation, maps, and other complex documents truly AI-ready.

- **Representation layers and modern data management**: AI-ready datasets should be managed across multiple representation layers, each optimised for specific technical and operational needs, employing bronze, silver, and gold categorisation as required to represent raw, cleaned, and curated data layers respectively. This layered approach helps ensure that data can be efficiently stored, accessed, and reused for different purposes—making it easier to support advanced AI capabilities while maintaining reliability and security.
    - At the storage layer, robust formats such as Parquet or Iceberg should be used for immutable logs and structured data, while CSV is best avoided for complex or large-scale AI capabilities.
    - The semantic layer requires that unstructured data such as text and images be accompanied by precomputed vector embeddings stored in a vector database, enabling advanced AI capabilities like semantic search and Retrieval Augmented Generation (RAG); supporting vector databases indices (such as Facebook AI Similarity Search, FAISS or Hierarchical Navigable Small Worlds, HNSW) is essential for scalable storage and retrieval of embeddings.
    - For the access layer, protocols like Model Context Protocol (MCP) or GraphQL facilitate agentic, single record access, while zero copy Arrow enables high throughput training access.
    - Finally, the privacy layer should employ techniques such as Differential Privacy or Synthetic PII substitution to maintain linguistic utility and protect citizen data.

This layered approach ensures that data is not only efficiently stored and accessed, but also semantically rich, privacy preserving, and truly AI-ready.

- **Dataflow orchestration**: Efficiently manage data movements and transformations across multiple platforms layers and systems.
- **Include diverse perspectives**: Reflect different user and stakeholder perspectives decision approaches to reduce the likelihood of bias.
- **Manage dependencies**: address the relationships and dependencies between various data sources, systems, and processes to ensure seamless integration and minimise potential points of failure

[8] Sam Trendall, *HMRC algorithm for automated council tax valuations offers 'greater consistency and quality'*, *PublicTechnology*, published 6 August 2025, accessed 15 January 2026, https://www.publictechnology.net/2025/08/06/economics-and-finance/hmrc-algorithm-for-automated-council-tax-valuations-offers-greater-consistency-and-quality/

- **Data interoperability**: API design for accessibility: RESTful APIs should follow Open API standards, use clear HTTP methods, and provide features like pagination, filtering, and rate limiting. For complex queries, use SFTP for volume and GraphQL offers flexibility and precision.
- **Model Context Protocol (MCP)** and **agentic workflows** represent a modern approach to data interoperability[9]. It allows connections to different tools and datasets using Agentic AI. MCP is an open standard that lets data providers share resources through MCP servers, which describe their capabilities in a machine-readable format. AI capabilities act as MCP clients, requesting data securely and efficiently. Agentic AI capabilities further automate these processes, handling data pipeline tasks. This combination ensures data is discoverable, accessible, and ready for AI, helping the public sector improve data quality and meet interoperability challenges with robust technical standards and automation.

Example: i.AI was involved in the analysis of complex structured, semi structured and unstructured data —demonstrating the diversity of data types that must be managed for AI readiness. To enable this, multiple ingestion pipelines, transformation steps, and privacy controls had to be orchestrated across systems, explicitly reflecting the need for dataflow orchestration, dependency management, and interoperability across heterogeneous platforms. In this case, there was a need to address Personally Identifiable Information (PII); however, removing PII risked reducing the usefulness of the resulting data. This highlights the importance of balancing data privacy with maintaining data utility when managing complex and diverse datasets.

Example: A DSIT team found managing public consultations challenging, as aligning privacy notices across multiple departments required more than a year before AI based analysis could begin. This illustrates the real-world dependency management and cross-organisational coordination required before data can be safely integrated and orchestrated for AI use. Additionally, projects included converting legislation and case law into structured, machine readable formats and leveraging computer vision to digitise planning documents and historical maps, thereby creating valuable geospatial datasets. These initiatives demonstrate end-to-end dataflow orchestration across unstructured and structured sources, and the need for interoperable interfaces and standards to make data discoverable, accessible, and usable across departments and AI capabilities.

Scale and performance

Generally, AI capabilities operate over large volumes of data and increasingly require near real time access for inference. AI-ready datasets must therefore be designed with scale and performance as core design considerations. AI training and AI inference are the next steps which would form a part of AI capabilities lifecycle.

[9] Haishan Fu, Aivin Solatorio, Olivier Dupriez & Craig Hammer, *From open data to AI-ready data: Building the foundations for responsible AI in development*, World Bank Blogs, published 21 July 2025, accessed 15 January 2026, https://blogs.worldbank.org/en/opendata/from-open-data-to-ai-ready-data--building-the-foundations-for-re

Essential capabilities include:

- **Handling massive data volume**: Be able to scale to accommodate rapidly growing data without performance degradation.
    - **Capture data locally**: For AI at scale, compute resources (e.g., GPUs) should be positioned as close as possible to the data. Moving petabytes of data across networks to feed models is one of the biggest performance bottlenecks.
    - **Clarify scale beyond database size**: Scale isn't just about how large the database is; it's about minimising data movement. Reducing unnecessary data transfers is critical for maintaining efficiency and throughput.
- **Providing real time processing**: Support timely insights and actions with real time or near real time data processing.
- **Optimising resource utilisation**: Look to maximise efficiency and minimise costs in data processing, storage and its lifecycle management. Keeping all big data in real time storage is financially expensive. Use tiered storage and caching strategies to maximise efficiency and minimise costs across the data lifecycle. If possible, implement Storage Tiering & Intelligent Caching:
    - **Hot Storage / Streaming**: Think of this as the "active workspace." It's where the model keeps the data it's currently learning from like the notes you're using right now to study. Example: in neural networks, current epoch.
    - **Warm Storage (S3/Object Store):** This is the "bookshelf." It holds the entire dataset that the model uses during training sessions. It's not in your hands right now, but you can grab it easily when needed.
    - **Cold Storage (Infrequent Access):** This is the "attic." It stores old logs or historical data that you might need later—say, within the next six months for retraining. Avoid putting anything here if you'll need it soon, because getting it back takes more effort.

Example: The Department for Transport's provision of real time traffic data[10] via APIs demonstrates good resource utilisation. By enabling continuous updates and high frequency access, these datasets support AI driven smart city, congestion management, and optimisation use cases. Crucially, these services are designed to operate at national scale, handling very high data volumes while minimising unnecessary data movement by exposing standardised, queryable interfaces close to the source systems. This allows models to consume live streams for inference while drawing on historical stores for training, aligning hot, warm, and cold data tiers to optimise performance and cost.

Practical recommendations for pillar one

- **Data planning**: Maintain versioned schemas and clear deprecation policies

[10] Highways England, *WebTRIS Traffic Flow API* (dataset), published 30 November 2020, last updated 5 August 2022, accessed 15 January 2026, https://findtransportdata.dft.gov.uk/dataset/webtris-traffic-flow-api

- **Data preparation**: Use modern, interoperable formats (TXT, CSV, Parquet, JSON, XML, GeoParquet).
- **Data provisioning**: Provide APIs (e.g. REST/OpenAPI, SFTP, GraphQL) with pagination, filtering, and rate limiting to aid access to data products.
- **Data provisioning**: Design for scalability: support bulk downloads and streaming where needed.

Pillar 2: Data and metadata quality

This pillar starts from the premise that organisations seeking to apply AI must first be able to clearly articulate what data they need, for what purpose, and at what level of granularity.

Teams should be supported to discover what datasets already exist (for example through data marketplaces, catalogues, or registries) and to engage early with data owners and specialists to define access requirements, sensitivities, and intended uses. This upfront clarity is critical not only for technical feasibility, but also for establishing lawful basis, scoping DPIAs, and setting appropriate governance controls.

Building on this, the pillar addresses the trustworthiness and interpretability of datasets covering three AI-ready data principles. AI capabilities rely on both data values and metadata to learn effectively and to produce outputs that can be explained, audited, and defended.

The three principles reinforce treating datasets as data products, operation of a Data Quality Action Plan and consistent metadata practices are.

Data quality

Data quality, including consistency and freshness – a robust data environment is essential for maintaining high standards of data quality and consistency across the organisation. A Data Quality Action Plan[11] (DQAP) should be developed, maintained and made available to users. Assigned owners of the DQAP should use a Prioritisation Framework[12] to assess which data quality issues are critical for resolution.

Key expectations include:

- **Preserve data integrity**: Maintain accuracy and reliability as data flows between disparate sources and systems.

[11] UK Government, *Data quality action plan: implementation guide*, published 11 November 2025, accessed 15 January 2026, https://www.gov.uk/government/publications/implementing-a-data-quality-action-plan/data-quality-action-plan-implementation-guide
[12] UK Government, *Data quality issues framework*, published 11 November 2025, accessed 15 January 2026, https://www.gov.uk/government/publications/implementing-a-data-quality-action-plan/data-quality-issues-framework

- **Standardise data formats**: Data quality approach requires consistency in the formats and representations while addressing the quality issues from source into the raw storage layer or during the data processing to enable seamless integration and interoperability.
- **Implement data validation mechanisms (observability):** Detect and correct anomalies to ensure data remains trustworthy for AI capabilities.
- **Data freshness**: Timely data delivery can be essential for effective AI, requiring low latency, real time pipelines. Example: Real-time for fraud; Daily/Weekly for reporting; Monthly for strategic planning. Techniques like change data capture and stream processing ensure fresh data from various sources. Continuous updates to repositories enable accurate and up to date AI predictions.
- **Data must be accurate**: Data must be accurate for AI capabilities to function reliably. AI capabilities can only learn from the information they are provided, so while datasets do not need to be fully complete, the data that is available should be correct and trustworthy. If information is inaccurate it may lead to biased outputs, nonsensical creations and a malfunctioning AI system. Measuring the data accuracy is quite complex in addressing all the challenges, but it can be efficiently supported by.
    - Profiling source data through Exploratory Data Analysis (EDA) to understand its characteristics, completeness, distribution, redundancy and shape.
    - Data Quality Monitoring is an automated and continuous process to monitor data quality by building dashboards for each data product, providing that view to business and technical stakeholders to proactively alert in case of any anomalies, this should also include data drift observability which are usually achieved by commercially off the shelf products integrating them to the datasets or building a custom observability pipelines across the datasets.
    - Data lineage and impact analysis to highlight the impact of data changes and trace the origin of data to prevent accidental modification of the data used by AI capabilities.

Data as strategic product

AI readiness requires datasets to be treated as strategic data products rather than static publications. A Data product includes data contracts, code assets, metadata, and related policies, aiming to provide long term business value to intended users through usable, valuable, searchable and feasible solutions.

Essential capabilities for AI readiness:

- Treat datasets as products with owners, interfaces, SLAs, and catalogue entries. This entails assigning ownership, discoverable, defining interfaces, committing to service levels, and aligning dataset management with the FAIR principles of Findability, Accessibility, Interoperability, and Reusability

Example: HMRC's tax datasets[13] illustrates this model through the provision of an API, documentation, update schedules, and support contacts, positioning the dataset as a reusable asset rather than a one-off release.

Business logic and context

Business and Context driven metadata needs to be captured not only from outcomes of business decisions but from sources, underlying logic and through effective business inclusion that could describe what the data is, how it's structured, who maintains it, and how it can be accessed. The GDS metadata schema for describing data assets exchanged between UK government organisations should be followed.

Essential capabilities for AI readiness:

- **Capture business logic** - AI-ready datasets must capture not only observable data values but also the business logic, decision rules, and constraints that underpin data collection and transformation. AI capabilities infer meaning from patterns in data, and where contextual information is absent, they risk learning correlations that are misleading or inappropriate for the intended use. Data lineage and impact analysis to highlight the impact of data changes and trace the origin of data to prevent accidental modification of the data used by AI capabilities
- **Context** - Organisations should therefore document the purpose for which data was collected, the operational or policy decisions it reflects, and the assumptions and trade-offs that shaped its transformation. This includes distinguishing between data intended for direct service delivery and data suitable for system wide analysis, research, or secondary use.

Example 1: NHS leaders emphasised that operational datasets commonly distinguish between information used for direct patient care and information that may be reused at an aggregated level for system wide analysis or research. Documentation clarifying aggregation thresholds, sensitivity, and acceptable secondary uses helps to prevent inappropriate AI deployment and supports compliance with clinical, ethical, and public expectations.

Example 2: National Archives mentioned that don't use subject-based taxonomy as the core as, try to maintain how records were organised by whoever was creating them, and you gain context by the records' relationship one to another Maintaining these original relationships ensures that the meaning, provenance, and constraints embedded in the data are not lost—an approach that is critical for AI use, where models must be grounded not only in content, but in the operational and institutional logic that shaped it.

[13] HM Revenue & Customs, *HMRC Developer Hub: API documentation* (online portal providing access to HMRC API specifications and integration guidance), accessed 15 January 2026, https://developer.service.hmrc.gov.uk/api-documentation

Practical recommendations for pillar two

- **Data preparation**: Operate a Data Quality Action Plan (DQAP) with metrics covering accuracy, completeness, consistency, timeliness, validity, uniqueness. A quality dashboard could be set up, to support trustworthiness
- **Data provisioning**: Publish rich metadata such as ownership, licensing, update schedule, quality status to support discoverability. Standards such as Dublin Core and DCAT serve as examples of best practice for context driven metadata.
- **Data provisioning**: Treat datasets as data products with clear contracts and SLAs.

Pillar 3: Organisation and infrastructure context

This pillar recognises that AI-ready data cannot be achieved through technical measures alone. Organisational commitment, clear roles, and sustainable infrastructure are essential to maintaining readiness over time. The three principles reinforce data governance, sharing & collaboration and adequate skills, documentation & guidelines.

Data governance

AI readiness starts with strong data governance. By establishing clear policies, ensuring data security, and maintaining transparency across the data lifecycle, organisations create a trusted foundation for ethical and scalable AI.

Essential capabilities for AI readiness:

- Assigning explicit **Role Based Access** Control (RBAC) roles for datasets meaning access to resources (like files, applications, or data) is determined by the user's role—such as Admin, Developer, or Viewer. Including responsibility for quality management, access decisions, incident response, and lifecycle management. Clear role definition enables faster resolution of issues and more effective governance.
- **Infrastructure monitoring** is critical as AI-ready data platforms must be operated sustainably, with visibility of performance, reliability, and cost.
- **Organisational transparency** and user engagement are essential for effective data reuse. Publishing change logs, provide clear support channels, and engage with users to understand evolving needs.
- **Global policies on privacy, compliance, and security** would need to be harmonised with domain specific policies and rigorously enforced. This includes the safeguarding of personally identifiable information (PII).

Example 1: Within the NHS, DPIAs for the federated data platform are routinely completed for new datasets and AI enabled uses, providing a documented basis for assessing privacy risks and safeguards

Example 2: Within DEFRA's Data Analytics and Science Hub (DASH) platform they demonstrate how access can be restricted at column, and table level, with full logging to support audit and assurance requirements, by enabling role-based and attribute-based access controls.

Information sharing and collaboration.

The most valuable AI insights often emerge from combining data across organisational boundaries. AI-ready datasets should therefore support governed information sharing and collaboration while enforcing purpose limitation and access controls.

Essential capabilities for AI readiness:

- Techniques such as breaking down data siloes and increasing interoperability, alongside prioritising common data standards to improve precise querying, purpose-bound access, and federated learning, enable AI capabilities to derive value from distributed datasets without requiring wider access across departments.

Example: NHS suggested that federated learning infrastructure enables secure data sharing for research by allowing models to be trained across multiple datasets without centralising patient records

Skills, documentation and guidelines

The effectiveness of AI-ready datasets depends not only on technical quality but also on the ability of users to understand and apply them appropriately. Departments must therefore invest in documentation, guidelines, and skills development alongside data publication.

Essential capabilities for AI readiness

- **Access to data stewards and subject matter experts** who understand the data is essential. Clear documentation, supported by a well-maintained knowledge repository should include business and technical glossaries, recorded source-to-target mappings, data processing and lineage documentation. Architectural and Business Decision Records management, Business Decision Records should be maintained to track evolving business needs, , whilst onboarding materials and accessible support channels reduce the risk of misuse and improve the quality of AI outcomes. In parallel, teams should be supported to develop skills in cloud data operations, privacy, and AI assurance.

Example: The Department for Education[14] provides clear documentation and support channels for school performance datasets, enabling responsible and effective reuse

Practical recommendations for pillar three

- **Data requirements**: Sufficient skills and documentation key to drive continuous improvement.

[14] Department for Education, *Data catalogue – Explore education statistics*, accessed 15 January 2026, https://explore-education-statistics.service.gov.uk/data-catalogue

- **Data planning**: Implement data governance, assign stewardship, quality, and incident response roles. This could be accomplished using a RACI matrix
- **Data acquisition**: Enable information sharing, data literacy and collaboration across departments.
- **Data preparation**: Monitor infrastructure for performance and cost.
- **Data provisioning**: Publish change logs and support channels.

Pillar 4: Legal, security and ethical compliance

This pillar ensures that AI-ready datasets are released and used in ways that are lawful, secure, and ethically defensible.

Given the sensitivity of many public sector datasets and the potential impact of AI driven decisions, this pillar is foundational to maintaining public trust.

- Departments must implement appropriate technical and organisational controls, including role-based access control, encryption of data at rest and in transit, and comprehensive audit logging. These controls should apply consistently across file-based access, APIs, and analytical environments.
- Legal compliance requires clear evidence of the lawful basis for data processing, particularly where datasets may be reused for AI training or inference. Departments should complete and maintain Data Protection Impact Assessments (DPIAs) where required, documenting risks, mitigations, and decision rationales.

Example: At the NHS, DPIAs are routinely completed for new datasets and AI enabled uses, providing a documented basis for assessing privacy risks and safeguards.

Ethical considerations extend beyond formal legal compliance. AI-ready datasets should include documentation on acceptable use, known biases, and representativeness limitations. This is particularly important for datasets used in high impact contexts such as healthcare or social policy.

Example: NHS research datasets often include explicit notes on representativeness and known biases, such as limitations observed in skin analytics models, enabling more responsible interpretation and use.

Security and compliance

Security and legal compliance are non-negotiable requirements for AI-ready data.

Essential capabilities for AI readiness:

- Organisations must classify data appropriately, apply encryption at rest and in transit, implement role-based access controls, log access and changes, and clearly evidence

legal basis and data minimisation. These measures are essential to meeting statutory obligations and maintaining public trust in AI enabled public services.

Example: At The National Archives, recent measures to strengthen data security in the department include the introduction of Secure by Design principles[15]. They deployed monitoring tools such as Wiz and Wiz Defend for repository and supply chain risk management, and instigated a renewed emphasis on immutable backups and recovery planning particularly in response to incidents like the British Library cyberattack in October 2023

Practical recommendations for pillar four.

- **Data requirements**: Document the legal basis for data collection and processing, especially where data is used for AI training and inference and ensure it aligns with the original purpose of data processing. Apply data minimisation with purpose limitation and minimisation controls in data pipelines, including the use of pseudonymisation where possible.
- **Data planning**: Conduct a Data Protection Impact Assessment (DPIA), with a structured assessment to identify & quantify data privacy risks, such as discrimination, reidentification, or confidentiality loss and broader GDPR-aligned privacy governance with Data Subject Access Request (DSAR), consent management, incident management, and data mapping.
- **Data provisioning**: Applying platform-level role-based access control, attribute level access controls, clear permission workflows and identity integration, encryption via file level encryption at rest and in transit with network encryption between systems, and audit logging for compliance proof and breach detection.
- **Data provisioning**: Publish acceptable use notes like acceptable use, access, and data policies within Data Governance principles and bias disclosures with a need to identify dataset representativeness, document known biases and their implications and these will help users will understand limitations, mitigate discriminatory outcomes and comply with emerging AI regulations.

[15] UK Government, *Secure by Design principles*, published 12 November 2025, accessed 15 January 2026, https://www.security.gov.uk/policy-and-guidance/secure-by-design/principles/

AI-ready data action plan

To inform an action plan for AI readiness, seven interviews were conducted across government departments and arm's length bodies, complemented by a horizon scan of 12 organisations. It is becoming evident that the UK public sector is no longer constrained by a lack of AI ideas, but by whether its data is fit for purpose, legally usable, trusted, and operationally sustainable. Participants described a shift from "Can we build an AI system?" to "Can we safely and repeatedly use our data to support AI at scale?"

1. Technical optimisation and infrastructure

What's needed:

- APIs, semantic access, and modern data formats to enable machine actionable design.
- Strategies for unstructured data, such as correspondence or legal text, including tagging and annotation.
- Interoperability and incremental adoption of standards.
- Continuous monitoring and retraining to manage operational drift.

Departmental insights:

- **NHS** highlighted legacy technology burdens: over 50 CRM systems and 190 authorisation services create interoperability challenges.
- **DEFRA** manages over 500 paper form services, showing under digitisation issues.
- **The National Archives** raised concerns about digitising historical records without metadata, creating governance and risk gaps.

Key challenges:

- Legacy systems and lack of cloud skills are hidden blockers.
- Unstructured data is the biggest opportunity but also the highest risk.
- Operational sustainability issues emerge post deployment e.g. bias, performance degradation or misuse.

Mitigations:

- Treat unstructured pipelines as assurance critical systems with rights metadata and human in the loop controls.
- Use AI to create structured surrogates while keeping originals immutable and decisions reversible.
- Design for continuous feedback loops, monitoring model behaviour and service impact.

- Embed drift management and model retraining into the data lifecycle.

2. Data foundations and metadata quality

What's needed:

- Consistent, well-structured data designed for machine use.
- Robust metadata and a single source of truth.
- Synthetic datasets for testing and accelerating readiness.
- ML-oriented metadata including bias notes, provenance and versioning, alongside traditional descriptors.

Departmental insights:

- **DWP** emphasised inconsistent data structures as a barrier to scaling AI such as their 'DWP Ask' query tool.
- **ODI** reinforced that many datasets are published for transparency, not machine use, and therefore naturally lack APIs and version control.

Key challenges:

- Fragmented, inconsistent data prevents reuse of successful AI patterns.
- Metadata gaps make linkage and reuse fragile.

Mitigations:

- Build repeatable minimum data standards for specific AI patterns e.g., policy Q&A or prediction.
- Treat data modelling and structure as core infrastructure, not simply a by-product.
- Introduce ML-oriented metadata to support bias detection and provenance tracking.

3. Organisational context and stakeholder engagement

What's needed:

- Ongoing communication to build trust in synthetic data, redaction, and bias testing.
- Co-design with policy owners and frontline professionals.
- Institutional incentives, authority, and sustained funding.
- Clear accountability models separating dataset, service, and AI system ownership.

Departmental insights:

- **NHS** stressed the need for maintaining clinical accountability even with AI-driven decisions.
- **Esynergy** warned that "accountability without authority fails", slowing coordinated change.

Key challenges:

- Accountability structures lag behind AI reality.
- Data owners often lack authority to enforce standards or investment.

Mitigations:

- Empower senior roles with combined data/AI leadership functions.
- Adopt operating models which recognise AI capabilities as ongoing services requiring monitoring and re authorisation.

4. Legal, security, and ethical compliance

What's needed:

- Clear governance for PII, hosting, and defensible redaction.
- Ownership, licensing, and lawful basis for computational use.
- Consistent interpretation of GDPR and related laws.
- Explicit risk appetite statements and ethics integration from ideation.

Departmental insights:

- **i.AI** flagged inconsistent GDPR interpretation across the public sector as a major blocker, and highlighted the public task[16] which permits specific lawful basis of tasks to be permitted in the public interest.
- **National Archives** stressed rights aware governance and service protection controls as non-negotiable.
- **Esynergy** noted lack of executive level risk appetite statements can default decisions to "no".

Key challenges:

- Legal interpretation and unclear ownership dominate decisions.
- Departments lack authority and accountability clarity.

[16] Information Commissioner's Office, *Public task (a guide to lawful basis under UK GDPR)*, published October 2022, accessed 15 January 2026, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/public-task/

Mitigations:

- Move from generic compliance to dataset level legal decisioning such as lawful basis and withdrawal rights.
- Publish senior level risk appetite statements to enable proportionate decisions.
- Embed governance and ethics expertise early in the process.
- Design service protection mechanisms such as rate limiting and controlled APIs as part of readiness.

AI readiness cannot be reduced to a single universal checklist. It must be evaluated in relation to each specific dataset, intended AI use case, and the organisational, legal, and operational context in which it will be applied. Nevertheless, based on collective cross-government evidence and practice, we can articulate the core dimensions that define AI readiness. The following section presents a structured self-assessment checklist to support organisations in assessing whether a dataset can responsibly sustain AI use over time.

Self-assessment checklist

To develop a mature data enterprise, we need to think beyond simply whether the data exists but to establish whether a given dataset can responsibly sustain AI capabilities use over time.

A readiness scoring framework (e.g. using a 4-point scale) could be applied across the four foundational pillars and model families to help assess dataset suitability for proof of concept to production grade AI use cases.

To support greater awareness of the components of AI-ready data development, this section provides an example checklist of questions that an AI development team could work through to understand they have appropriate systems and processes in place.

1. Purpose and legitimacy

Question	Response
I know exactly what ML or generative AI task this data supports (e.g. prediction, classification, retrieval, summarisation, or decision support).	
I know whether outputs affect individuals or only cover population level insight.	
I know which uses are explicitly out of scope such as no automated decisions.	
There is a named product/policy owner who approves AI uses.	

2. Legal, rights, and governance readiness

A credible self-assessment must rigorously test, for each priority dataset, whether legal status, rights, and governance arrangements can be clearly evidenced. Across interviews, organisations consistently stressed that dataset level legal clarity is a foundational prerequisite for any claim of AI readiness.

Question	Response
I know if this data can be used for training, fine tuning, inference or evaluation.	
Sensitive fields are clearly flagged (e.g. PII, confidential, special category).	

I know where I am allowed to work with this data (e.g. secure environment, APIs, offline).	
I know what happens if an AI use case changes.	

3. Metadata fitness

ODI and DWP both underlined that many datasets may be “open” or “available” yet fundamentally unfit for AI without targeted remediation. Key actions:

Question	**Response**
I understand what each table and field actually means.	
Event time, update frequency, and history are clear.	
Schema changes and versions are tracked.	
Provenance is known i.e. where the data comes from and how it is produced.	
Known data quality issues and biases are documented.	
Stable IDs exist (or are explicitly missing).	

4. Unstructured data pipelines

Both the Open Data Institute (ODI) and the Department for Work and Pensions (DWP) underlined that the availability or openness of a dataset does not equate to AI readiness; many datasets remain fundamentally unfit for AI use without targeted remediation. Key actions include:

Question	**Response**
Data is machine readable or has OCR / extraction pipelines.	
Rights, redaction, and sensitivity travel with extracted content.	
I can trace embeddings, chunks, or features back to source files.	
There are human review steps for training data and outputs.	
There is a defined pipeline to ingest documents, images, audio, or free text.	

5. Operating model and sustainability

Sustained AI use must be explicitly assessed, including how drift, bias, misuse, and service impacts are governed. NHS and The National Archives warned that without this, early success can translate into long term risk.

Question	Response
I know who owns data quality, drift, and bias aware monitoring.	
There are pipelines to detect breaks, drift, and anomalies.	
I know how data or model issues are fixed and re approved.	
Training datasets can be reconstructed exactly.	
There is a process if AI causes service, cost, or trust issues.	

Conclusion

The UK public sector stands at a pivotal moment in the evolution of artificial intelligence in government. While the availability of advanced models and platforms continues to accelerate, these guidelines demonstrates that the true constraint on responsible, scalable AI adoption is not algorithms, but data: its quality, structure, governance, and legitimacy. Releasing datasets for AI use is therefore not a technical publishing exercise, but a strategic act of public stewardship.

This document has set out practical, UK specific guidance for preparing and releasing datasets that are genuinely AI-ready datasets that are technically fit for modern AI capabilities, intelligible beyond their original operational purpose, and governed in ways that are lawful, ethical, and worthy of public trust. Drawing on existing statutory and policy frameworks, and grounded in cross government interview insights, it reframes data release from a one-off compliance activity into the continuous management of datasets as strategic public assets.

Across the four foundational pillars technical optimisation; data and metadata quality; organisational and infrastructure context; and legal, security, and ethical compliance the guidelines makes clear that AI readiness is inherently socio technical. Infrastructure modernisation, metadata fitness, and unstructured data pipelines are essential, but insufficient without clear accountability, sustained skills, and explicit legal and ethical decisioning at dataset level. Departments consistently reported that the principal barriers to AI are not a lack of ideas or tools, but uncertainty around rights, inconsistent governance, legacy data foundations, and the absence of operating models capable of sustaining AI capabilities over time.

A central conclusion from the interviews is that AI readiness cannot be reduced to a universal checklist. Readiness must be assessed in relation to specific datasets, specific AI use cases, and specific organisational contexts. The self-assessment framework included in these guidelines reflects this shift: from “Do we have data?” to “Can this dataset responsibly sustain AI use across its full lifecycle?” This encompasses not only model development, but ongoing monitoring, drift management, re authorisation, and the governance of service impacts on cost, delivery, and public confidence.

Critically, these guidelines position AI-ready data as a managed, end to end capability. Departments that are most advanced are those that treat datasets as products with named owners, explicit legal bases, quality obligations, and user support; that embed ethics, security, and risk appetite at the outset; and that recognise AI capabilities as continuing public services rather than finite technology deployments. Where this mindset is absent, early successes risk creating long term operational, legal, and reputational exposure.

By adopting the principles, pillars, and action practices set out in this document, UK government organisations can move from opportunistic AI experimentation to a sustainable national capability for responsible AI. This will enable departments not only to release data more effectively, but to do so in a way that strengthens interoperability, accelerates safe

innovation, and reinforces public trust. In doing so, HMG has the opportunity to establish global leadership not merely in AI application, but in the stewardship of public data as critical national infrastructure for the age of intelligent systems.

Appendix A – Methodology

To ensure this paper incorporated relevant actionable content, the following methodology was undertaken:

- Literature review covering international government standards and recognised frameworks from leading research bodies, including the Open Data Institute's (ODI) *Assessing Data Readiness* paper and the World Bank's *From Open Data to AI-Ready Data* reports.
- Interviews with domain experts, including senior digital leaders and data policy advisors from the public bodies and government departments and technology consultants from specialist AI and data firms. These structured interviews captured lessons learned and best practices for implementing AI initiatives in a public-sector context.
- Data and technology providers were consulted and reviewed the initial draft publication, helping to ensuring guidance was grounded in proven practices to deliver credible, consistent, and actionable outcomes for government data readiness.

Appendix B – Acknowledgements

This document has been developed with the invaluable support of multiple stakeholders across departments, external organisations, suppliers, and partners.

We would like to acknowledge the contributions from the following. Their input—whether through providing currently relevant information or reviewing these guidelines—has been instrumental in shaping this work.

- Department for Work and Pensions (DWP)
- National Health Service (NHS)
- Incubator for Artificial Intelligence (i.AI)
- The National Archives
- Open Data Institute (ODI)
- Public Sector AI Adoption Unit
- Esynergy
- Cognizant
- AWS
- Microsoft
- Databricks
- Faculty

Appendix C - Insights on AI readiness of government datasets

To capture a realistic picture of good practice and expert advice, a number of interviews with departments and expert organisations were conducted. Key insights from these interviews have informed the development of these guidelines. Some highlights from these interviews are given below:

Department for Work and Pensions (DWP)

The Department for Work and Pensions (DWP) manage some of the UK government's largest and most sensitive operational datasets, spanning welfare, employment services, pensions, and fraud and error. These data assets are characterised by high volume, continuous change, and extensive personally identifiable information (PII). DWP provided insights from their recent experience developing the Generative AI tools DWP ASK, an internal policy chatbot which used Retrieval Augmented Generation (RAG) and GAIL, a tool that assists learning designers to create new content.

- **Frame AI readiness** primarily as a data foundation and governance challenge, rather than a modelling or tooling issue.
- **Data consistency & machine actionable design**: DWP emphasises that inconsistent data quality and structure are major barriers to scaling AI. Success depends on designing data for machine use, not just human readability.
- **Synthetic data:** DWP uses synthetic data to accelerate AI readiness, especially when access to real data is restricted.
- **Governance & PII**: Governance, hosting, and personally identifiable information (PII) constraints are fundamental. Secure redaction and hosting are essential.
- **Metadata & single source of truth**: DWP invests in creating robust metadata and a single, well governed source of truth to speed up future AI projects.
- **Expectation management**: Managing stakeholder expectations and building trust in synthetic data and bias testing is as important as technical work.
- **Reusable patterns**: DWP notes that reusable AI patterns depend on mature data foundations.

National Health Service (NHS)

The National Health Service (NHS) oversees one of the most complex data ecosystems in the UK public sector, spanning clinical care, population health, operational management, and

biomedical research. Its datasets range from highly sensitive individual level health records to large scale population and service datasets, distributed across diverse providers, systems, and standards. As a result, NHS perspectives on AI readiness are shaped by clinical safety obligations, professional accountability structures, and the central importance of public trust. Interviewees consistently framed AI readiness not as a uniform technical state, but as a context dependent condition, varying significantly by purpose, scale, and proximity to direct patient care.

- **Purpose, scale & acceptability**: NHS finds that the acceptability of AI use depends on the purpose and scale population level uses are less contentious than individual level applications.
- **Bias as clinical safety**: Bias and representativeness in training data are clinical safety issues. NHS stresses the need for demographic coverage and ongoing monitoring.
- **Accountability for autonomous AI:** NHS highlights challenge in deploying autonomous AI due to deeply embedded professional responsibility in clinical practice.
- **Model drift**: NHS identifies model drift as an operational risk, requiring continuous monitoring and feedback loops.
- **Public trust & opt out**: NHS stresses the importance of public trust, clear consent models, and respect for opt out boundaries.
- **Fragmented standards**: NHS operates across fragmented data standards, requiring interoperability and provenance controls.
- **Tiered AI opportunities**: NHS classifies AI opportunities into corporate automation, administrative workflows, and clinical decision support, each with escalating data readiness demands.
- **Contextual “good data”:** NHS asserts that “good data” is contextual and must be assessed for each use case.

Incubator for Artificial Intelligence (i.AI)

The Incubator for Artificial Intelligence (i.AI) operates at the intersection of policy, delivery, and experimentation across the UK public sector, supporting departments to translate emerging AI opportunities into deployable services. Unlike operational departments that manage a bounded set of mission datasets, i.AI engages with highly heterogeneous data landscapes, spanning legacy systems, policy repositories, service data, and large volumes of unstructured content. As a result, I. Ai’s perspective on AI readiness is shaped by cross government legal interpretation, institutional incentives, and the practical realities of modernising fragmented data estates. Interviewees consistently framed AI readiness less as a property of individual datasets, and more as a system level condition combining legal clarity, technical foundations, and trusted relationships to enable responsible data use across organisational boundaries.

- **Ownership & licensing**: i.AI finds that clarity over data ownership, access rights, and licensing is foundational for AI readiness.

- **Law interpretation**: Inconsistent interpretations of data protection law (e.g., GDPR) are significant barriers, requiring shared institutional memory.
- **Unstructured data**: i.AI identifies unstructured data as the frontier for public sector AI, with governance often lagging behind technical capability.
- **AI created data**: i.AI notes that AI itself often creates AI-ready data by converting legacy assets into structured formats.
- **Trust & transparency**: i.AI stresses that trust, openness, and transparency are integral to data readiness.
- **Stakeholder engagement**: Building trusted relationships with policy owners and frontline professionals improves data quality and willingness to share.
- **Incentives & authority**: Institutional incentives and authority are crucial for enabling AI readiness.
- **Technical foundations**: APIs, semantic access, and modern data formats are necessary; legacy systems and limited cloud skills are blockers.
- **Federated knowledge infrastructure**: i.AI envisions a federated, semantic, and curated national knowledge infrastructure.

The National Archives

The National Archives serves as the UK government's guardian of the official record, stewarding an exceptionally heterogeneous corpus of digital and physical assets spanning legal, administrative, and historical domains. Its data estate encompasses highly structured registers, complex legal records, digitised collections, and large volumes of born digital and unstructured material. As a result, TNA approaches AI readiness through the combined lenses of archival integrity, rights management, long term preservation, and public access. Interviewees consistently emphasised that AI readiness in an archival context is not a single technical threshold, but a context specific transformation challenge, where provenance, representation choices, and preservation grade metadata are as critical as machine learning performance.

- **Heterogeneous data landscape**: The National Archives manages diverse data types, requiring attention to provenance, rights, and context.
- **AI agents & demand shocks:** The rise of AI agents and bots creates new operational challenges, requiring robust technical controls.
- **Rights & remedy**: Active management of rights, access, and remedy is essential, especially for sensitive records.
- **Representation engineering**: Simplifying structured information for AI consumption is sometimes necessary, even at the cost of semantic precision.
- **Deep modelling for legal data**: Mature modelling involves deep versioning and graph-based relationships.

- **Preservation metadata**: Preservation grade metadata standards enable access rights and archival context.
- **Security & resilience**: Security and resilience are prerequisites for public data services in the AI era.
- **Narrow task evaluation**: Evaluation of AI capabilities are most effective in narrow, measurable domains.
- **Context specific readiness**: AI readiness is not binary; each dataset or service may require specific transformation.

Open Data Institute (ODI)

The Open Data Institute (ODI) operates as a cross-sector authority on data stewardship, standards, and data enabled innovation, working with government, local authorities, and industry to improve the foundations on which data driven systems are built. Its perspective on AI readiness is therefore rooted less in individual use cases and more in the systemic conditions that enable data to be discoverable, governable, and reusable across organisational boundaries. Interviewees consistently framed AI readiness as a direct extension of mature data practice encompassing interoperability, metadata quality, procurement design, and licensing clarity rather than as a discrete AI capability. As a result, ODI's insights focus on the structural and institutional barriers that prevent otherwise "open" or "available" datasets from becoming genuinely usable for machine learning and generative AI.

- **Data foundations**: ODI sees AI readiness as an extension of good data foundations (findability, accessibility, interoperability, reusability).
- **Local gaps & procurement**: Readiness gaps often stem from procurement and interoperability failures, not technical AI issues.
- **Technical scenarios**: ODI identifies three scenarios information retrieval, classical machine learning, and Generative AI each with distinct requirements.
- **Technical blockers**: Common blockers include reporting oriented formats, sparse metadata, lack of APIs, and weak version control.
- **Governance**: Governance is a first-class dimension, with many blockers rooted in procurement and stewardship.
- **ML quality vs traditional quality**: Machine learning requires checks for bias and representativeness, not just traditional data quality.
- **Licensing constraints**: Licensing remains an unresolved constraint for generative AI.
- **Metadata as barrier**: Metadata quality is a critical barrier to AI readiness.
- **Archival content**: Archival content needs a maturity pathway, including digitisation, metadata capture, and governance.

Esynergy

Esynergy operates across central and local government as a delivery and transformation partner, working with departments to move digital, data, and AI initiatives from concept into operational services. Its perspective on AI readiness is therefore shaped by repeated exposure to the practical gap between proof-of-concept activity and sustained deployment. Interviewees consistently framed AI readiness not as a technical property of datasets alone, but as an end-to-end organisational capability, encompassing data infrastructure, delivery models, governance authority, and executive risk ownership. From this vantage point, AI readiness is fundamentally about whether institutions can reliably convert existing data assets into trusted, operational systems under real political, ethical, and delivery constraints.

- **Infrastructure deficit**: Esynergy finds that many public sector organisations lack strong data infrastructure and understanding of data pipelines.
- **Deployment gap**: The gap between proof of concept and operational deployment is often a data and delivery problem.
- **Clarify use case**: Clarifying the AI use case is essential before specifying data readiness.
- **Start with existing data**: Readiness should begin with what data already exists, not what is wished for.
- **Ethics & privacy**: Ethics, privacy, and societal trust must be integrated from the outset.
- **Risk appetite**: Clear, executive approved risk appetite statements enable consistent decisions.
- **Accountability & authority**: Governance models must align authority, funding, and delivery capability.
- **Realistic standards:** Standards and interoperability should be approached incrementally and realistically.
- **Unstructured data strategy**: For unstructured data, tagging, annotation, and access controls are as important as storage.

Faculty

Faculty operates as AI transformation partner in the government sector to effectively deploy AI by transforming raw, complex data into high-quality, AI-ready assets through advanced analytics, domain-specific data engineering, responsible data standards, and proven delivery across public sector, defence, healthcare, and national security. From their experience of working in the public sector the data readiness they have eventually addressed for the use cases developed at few departments like London Councils, DfE is to ensure data is clean, structured, labelled, governed, and operationally ready for high-impact AI use cases.

They explained that while creating AI-ready dataset especially for the DfE Content Store[17] program where the core job is to take policies, OFSTEAD data, student hand-writing submissions, PDF, word and spreadsheets and make it AI-ready so that EdTech tools can consume it. Some of the few challenges were governance and ensuring information remains aligned with the official source, metadata, data quality and privacy rules.

- **Technical Challenges**: If DfE updates policy, the system must know otherwise it risks being out of sync, lacking trust and accountability. To resolve the Content Store becomes a middle layer between the public source and AI consumers until the policies are published in a machine-readable form
- **Metadata**: While working with tabular datasets (outside Content Store), metadata is often minimal or missing, complicating AI readiness, department is building a layer to address the metadata management of structured datasets.
- **Dual Interface Design**: Content store had to support technical users where they needed API access and non-technical users like product owners need web interfaces
- **AI modes usage**: Used frontier LLMs (e.g., GPT-4) for summarisation, trained custom transformer models to classify copyright pages, performed manual labelling where no training data existed

[17] Faculty, *Faculty and expert education organisations to lead programme to put safe, impactful AI in the classroom*, published 3 October 2024, accessed 15 January 2026, https://faculty.ai/insights/articles/faculty-ai-and-expert-education-organisations-to-lead-programme-to-put-safe-impactful-ai-in-the-classroom

Appendix D - Examples from the UK and overseas

The UK government is not alone in its efforts to make its data AI-ready. Several other countries and organisations are developing their own frameworks and standards to support AI-ready data.

HMG must learn from these approaches while charting its own path aligned with national priorities and the opportunity that a National Data Library[18] could provide.

Examples include:

- **World Bank**: Leading the charge in making development data AI-ready[19] through comprehensive standards, MCP implementation, and international partnerships
- **Singapore**: Singapore's Model AI Governance Framework[20] and AI Verify testing toolkit consisting of 11 AI ethics principles provide readily implementable guidelines for private sector organisations addressing key ethical and governance issues when deploying AI solutions
- **Australia**: Australia's Technical Standard[21] for government AI use establishes consistent requirements and best practices for design, development, deployment, monitoring, and decommissioning of AI capabilities, using a reference AI lifecycle model ensuring holistic coverage
- **Genomics England** ensures its genomic data[22] is AI-ready for complex variant analysis through a robust data infrastructure, adherence to FAIR principles, and strategic partnerships. It maintains one of the world's largest genomic databases – the National Genomic Research Library (~50 PB of whole genome and clinical data) – within a secure, cloud based Trusted Research Environment. All genomic sequences and variants are stored in standardised formats (e.g. CRAM/BAM for sequencing reads and VCF for variant calls) and linked to rich clinical metadata using common ontologies like the Human Phenotype Ontology, ensuring interoperability and machine readiness. Genomics England's approach aligns with FAIR principles: data is highly findable and accessible to approved researchers via a governed research portal (over 1,500

[18] Government Office for Science, *National Data Library*, published 29 July 2025, accessed 15 January 2026, https://www.gov.uk/government/publications/national-data-library/national-data-library
[19] World Bank, *AI for Data – Data for AI* (World Bank Development Data Group programme overview), published 2025, accessed 15 January 2026, https://www.worldbank.org/en/about/unit/unit-dec/dev/ai-for-data
[20] Personal Data Protection Commission Singapore, *Model AI Governance Framework*, second edition published 21 January 2020, accessed 15 January 2026, https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework
[21] Australian Government Digital Transformation Agency, *Technical standard for government's use of artificial intelligence*, last updated 22 August 2025, accessed 15 January 2026, https://www.digital.gov.au/policy/ai/AI-technical-standard
[22] Genomics England, *Genomics 101: What is the National Genomic Research Library?*, published 20 May 2024, accessed 15 January 2026, https://www.genomicsengland.co.uk/blog/genomics-101-what-is-the-national-genomic-research-library

researchers securely analyse de identified data), interoperable through global standards (adopting GA4GH frameworks such as CRAM and the htsget API), and well documented and curated for reusability.

Appendix E - Other reference material

This document should be read alongside existing government frameworks and standards, which together provide the statutory, ethical, technical, and operational foundations that underpin responsible AI-ready data release. These resources guide both data & AI specialists and policy officials, clarifying when and how each should be used, helping help departments integrate AI-ready dataset practices within wider governance, transparency, ethics, and compliance requirements.

Core frameworks, standards, sector specific guidance aligned to the core pillars:

Technical & Legal

- UK GDPR[23] and ICO guidelines – set the statutory foundation for lawful data processing, including lawful basis, data minimisation, rights, and accountability.
- Understanding AI Capability – provides an accessible explanation of what AI systems can and cannot do, helping policy makers assess suitability, risks, and realistic expectations.
- The Government Data Quality Framework[24] – a framework describing how to understand, measure, document, and improve data quality throughout the data lifecycle.

Ethical & Operational

- Data and AI Ethics Framework[25] – a set of principles and activities guiding responsible, safe, and proportionate use of data and AI across government.
- Algorithmic Transparency Recording Standard (ATRS) Hub[26] – a standardised approach for publishing information about algorithmic tools and AI capabilities to promote transparency and public trust.
- NHS England's 'Data Saves Lives'[27] – sector-specific guidance illustrating how high-quality, well-governed data underpins safe and effective digital and AI services.

[23] Information Commissioner's Office, *The UK GDPR* (guidance on UK General Data Protection Regulation), last updated December 2025, accessed 15 January 2026, https://ico.org.uk/for-organisations/data-protection-and-the-eu/data-protection-and-the-eu-in-detail/the-uk-gdpr/
[24] UK Government, *The Government Data Quality Framework*, published 29 January 2025, accessed 15 January 2026, https://www.gov.uk/data-ethics-guidance/the-government-data-quality-framework
[25] UK Government, *Data and AI Ethics Framework*, updated 18 December 2025, accessed 15 January 2026, https://www.gov.uk/government/publications/data-ethics-framework/data-and-ai-ethics-framework
[26] UK Government Digital Service, *Algorithmic Transparency Recording Standard Hub*, published 5 January 2023, last updated 8 May 2025, accessed 15 January 2026, https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub
[27] Department of Health and Social Care, *Data saves lives: reshaping health and social care with data*, published 13 June 2022, accessed 15 January 2026, https://data.parliament.uk/DepositedPapers/Files/DEP2022-0485/Data_saves_lives_Reshaping_health_and_social_care_with_data.pdf

Glossary

Term	Definition
AI Opportunities Action Plan	A UK government plan (published January 2025) setting out priorities for accelerating AI adoption and value, including the need for accessible and trustworthy data.
Data Product	A dataset managed as a strategic asset with defined ownership, interfaces, quality standards, documentation, and service expectations, rather than a one-off publication.
Technical Optimisation	The design of data structures, formats, APIs, and infrastructure to support scalable, interoperable, and high-performance AI capabilities.
Data & Metadata Quality	The completeness, accuracy, consistency, freshness, and documentation of data and metadata required for AI capabilities to be interpretable, auditable, and trustworthy.
Business Logic and Context	Documentation of the purpose, decision rules, assumptions, and operational constraints that explain what data represents and how it should be used.
Data Quality Action Plan (DQAP)	A formal, owned plan defining how data quality is measured, monitored, prioritised, and improved over time.
Unstructured Data Pipeline	The governed processes and tools used to ingest, digitise, extract, annotate, and manage documents, images, audio, and free text for AI use.
Interoperability	The ability of datasets and systems to work together through shared standards, formats, and interfaces, enabling reuse across organisations and AI capabilities.
Data Governance	The roles, policies, and controls that define accountability for data quality, access, security, lifecycle management, and incident response.
Legal Basis and DPIA	The documented lawful grounds and impact assessments that justify data use, identify risks, and evidence compliance with data protection and related legislation.
Operational Sustainability	The capability to maintain AI enabled data use over time, including monitoring drift, bias, misuse, performance, and service impacts.
AI Readiness Self-Assessment	A structured evaluation of whether a dataset can responsibly support specific AI use cases, covering technical fitness, legal clarity, metadata, pipelines, and governance.

This publication is available from: https://www.gov.uk/government/publications/making-government-datasets-ready-for-ai

If you need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.