Data: the Lifeblood of UK AI Startups

July 2025

STARTUP
COALITION

Author

Vinous Ali
Deputy Executive Director
Startup Coalition

About Startup Coalition

The Startup Coalition is an independent advocacy group that serves as the policy voice for the UK's technology-led startups and scale-ups. We were founded in 2010 by Mike Butcher, Editor-at-Large of TechCrunch, and Jeff Lynn, Chairman and Co-Founder of online investment platform Seedrs. We fight for a policy environment that enables early-stage British tech companies to grow, scale, and compete globally.

Our network includes over 4000 startups, scale-ups, and investors. We have been instrumental in building proactive coalitions of businesses and investors on issues integral to the health of the UK's startup ecosystem. Our work has seen many successes, from establishing the Future Fund to creating the Scale-up visa.

Acknowledgements

We are grateful to all the startup founders who contributed to this report and supported us by providing their time, insight and expertise.

Introduction

The Prime Minister himself described AI as “the defining opportunity of our generation”[1] with the Government’s own sources citing that AI can boost the UK’s economy by 1.5 percentage points in productivity and be worth up to an average of £47 billion to the UK each year over a decade.[2] However, the window to capitalise on this opportunity is closing fast, and the UK must act swiftly and decisively to stay ahead.

It is generally accepted that the progress and capability of AI stands on three legs: the ability to access and use data, compute and talent. While all three are indispensable, data plays a foundational and highly critical role in the development of AI. Without the ability to access and use data, AI models cannot learn, identify patterns, make predictions or improve their performance.

Startups, especially, rely on publicly available datasets to innovate. They often aren’t able to leverage existing proprietary data nor have the resources to engage in data partnerships. Therefore, democratising access to usable data levels the playing field helping startups to compete with incumbents.

But where is this data coming from?

Much has been written about models trained on data that is publicly available online, but what is clear is that today, much of the world’s data, particularly high-quality datasets, are unstructured, inaccessible or unusable to researchers and developers alike. The National Data Library (NDL) represents a unique opportunity for the UK to unlock siloed public sector datasets - and those that sit in arms length bodies (ALBs) - to give UK startups and researchers a competitive edge in developing the next generation of AI systems.

At Startup Coalition, we believe the National Data Library could supercharge discovery and support mission-led Government by finding new pathways to solving some of the biggest challenges facing society today.

But with so much potential data to draw upon, sensitivities at every turn and IP and ownership issues, success is not guaranteed. Described as a ‘ten-year’ or ‘two-term’ project, it is clear that the Government is in it for the long haul. However, quick wins could put the NDL on the map and encourage greater engagement from even the most sceptical.

What follows is a roadmap for the National Data Library outlining three actionable workstreams to establish the NDL as a valuable interlocutor in the journey to AI leadership.

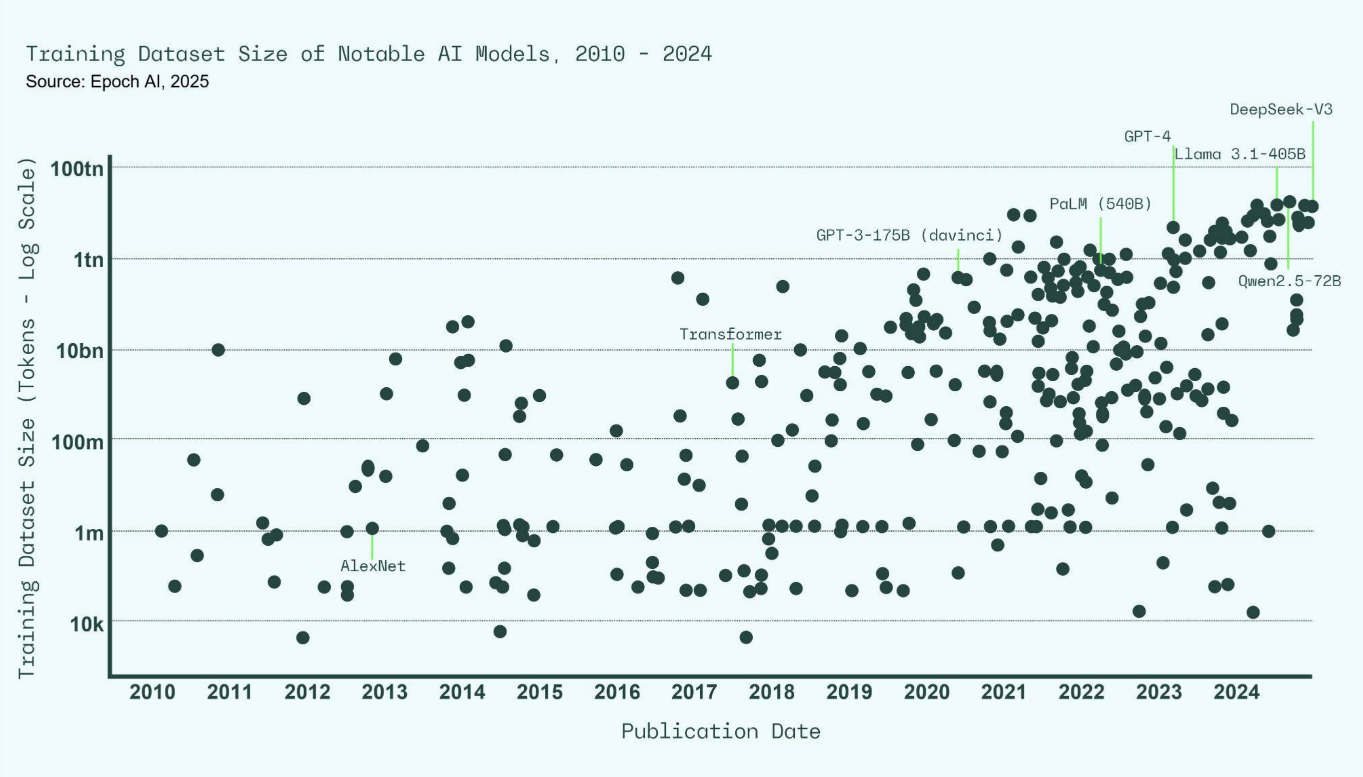
[1] https://www.theguardian.com/commentisfree/2025/jan/19/cringing-before-the-tech-giants-is-no-way-to-make-britain-an-ai-superpower#:~:text=lifetime%20in%20a%20speech%20delivered,%E2%80%9D

[2] https://www.gov.uk/government/news/prime-minister-sets-out-blueprint-to-turbocharge-ai#:~:text=Backing%20AI%20to%20the%20hilt,each%20year%20over%20a%20decade

The Critical Role of Data

Artificial Intelligence fundamentally depends on data. Classic machine learning techniques refine large datasets, and modern large language models learn by processing vast amounts of information to identify patterns, correlations and concepts across datasets. Scale as well as variety is key, directly impacting a model's performance, accuracy, and ability to generalise to new, unseen data. Neglecting either can lead to suboptimal or even failed AI deployments.

While frontier models, predominantly trained outside of the UK, have trended upwards in needing more training data over time. Therefore, the UK has an opportunity to seize on the opportunity higher up the stack, leveraging the pre-trained models and fine-tuning them.


Training Dataset Size of Notable AI Models, 2010 - 2024

Source: Epoch AI, 2025

Graph from Stanford AI Index 2025[3]

The UK is well positioned to accrue value at this layer in the stack given the following strengths:

1. **A robust R&D ecosystem:** The UK boasts world-class universities such as Oxford, Cambridge, and Imperial College London, which are consistently at the forefront of AI research. This academic excellence creates a strong pipeline of highly skilled AI talent. Institutions like the Alan Turing Institute further bolster national AI capabilities through research and collaboration.
2. **Government commitment and investment:** One of the first acts of this new Labour government was to commission what became the AI Opportunities Action Plan. This has been backed by £2bn in cash in the Comprehensive Spending Review with money going to build out AI infrastructure (including increasing compute capacity at least 20x by 2030) and investing in AI

[3] https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf

skills. These strategic investments create a fertile ground for fine-tuning large models, supporting deep research and pave the way for the development of large language models here in the UK in the future.

3. **Access to high-quality diverse data:** In theory, the UK has access to a vast and deep pool of data that could be unlocked in key sectors of the economy including financial services and health. The NDL has a crucial role to play in helping to make this a reality.
4. **A pro-innovation regulatory approach:** The UK has generally adopted a pro-innovation, principles-based approach to AI regulation, aiming to foster responsible innovation while balancing opportunity and risk. This approach, while evolving, seeks to avoid overly prescriptive regulations that could stifle development.

In essence, the success of AI progress in the UK requires both access to vast datasets and the ability to draw on context-specific, relevant data. This convergence reflects a maturing AI ecosystem: one where large general models and smaller, domain-optimised systems coexist. Where each one demonstrates a different demand on data availability and infrastructure.

Training is paramount for developing tailored predictive models or intelligent analysis systems. For instance, AI used in financial services, medical imaging, or e-commerce needs both very specific data from its particular field and broader data to understand fundamental human communication, writing and mathematics. This foundational knowledge is increasingly vital with the emergence of prompt engineering, chatbots and advanced techniques for retrieving and interpreting information.

Yet, a key obstacle for UK AI startups is the difficulty in accessing this crucial data, founders we spoke to reported that:

1. **There's not enough detail or granularity which is necessary for today's analytics.**
   To make high-quality analytical models, startups need record-level or highly detailed data to train. In order to compete with international competitors, startups want granular data that enables them to micro-target the needs of their customers and create micro-solutions that concretely tackle climate change, traffic. and more.

2. **The data is out-of-date too quickly.**
   Founders struggle when data, especially public sector data, is not updated frequently or has gaps in its time series. For many applications, consistent, real-time data access is a must. Without it, predictive analytics and services can't keep up with instantaneous changes, trends, or emergency situations. In turn, making their services delayed, unreliable, and without true optimisation.

3. **Founders have a tough time finding data in the first place.**
   Too often, Government data and "open" private sector data are only partially accessible. Startups lack the resources to navigate deep catalogues, restrictive licensing, or complex approval processes to access data. What they want is an easy way to access secure and privacy-compliant data quickly.

Founders worry that their access to relevant data (or lack thereof) could make or break the products they're trying to build, and they are fearful of wasting valuable time and resources on obtaining datasets that turn out to be unreliable. This sort of inaccessible data directly limits AI founders' ability to innovate.

Case Study: Soil Benchmark

Many ClimateTechs rely on data from the Geospatial Commission to track weather patterns, create predictive algorithms and more. But often, when they finally get their hands on this data, they find that it is outdated and incomplete, rendering it redundant. Unless we fix this, the founders' ability to innovate will continue to be stunted.

Soil Benchmark is a prime example. The startup helps farmers and landowners to make better decisions using high-resolution soil insights. This benefits everything from crop planning to reducing carbon-intensive fertiliser use, maximising the results for both agriculture and tackling climate change. But to do so, it needs access to foundational public datasets like the Cranfield soil maps (now known as LandIS). But these remain locked behind unnecessary restrictions despite being originally funded by the taxpayer.

On top of this, fragmented and un-stable subsidy schemes and limited data infrastructure in agencies like the Rural Payments Agency make it harder for startups in the same spaces as Soil Benchmark to develop their products. Unlocking datasets like LandIS, investing in modern API access, and harmonising how schemes collect data in the agri-economy would make it easier for startups to innovate in the agri-tech and climate spaces.

The UK's Unique Position

Few other nations hold the breadth of high-quality public-sector data that the UK has amassed, for example:

- Longitudinal health records in the NHS, with uses in medical AI research
- Detailed student data from our education system can provide details on literacy and skills
- Crime statistics, immigration records, and justice system data, usually found at the Home Office
- Extensive environmental and geospatial data, like climate data, land use, and satellite imagery

The list goes on and on.

Much of this is in part due to the UK's pioneering role as a global leader in open data and data transparency. In the 2010s, the UK led the global charge on Open Data, launching government websites and adopting and creating licenses that made it easier for startups to access public domain datasets. This resulted in apps like Citymapper, that used publicly held and open data to unlock community-wide benefits, to become a success overnight.

This was the beginning of standards and metadata practices, and the UK's reputation for pushing forward a digital public sector. We see this in the rapid rise of data-driven solutions in the UK, including:

- **Mastodon C,** which in collaboration with Open Healthcare UK, uses open NHS prescription data to uncover massive cost savings for the health service. By analysing publicly released data on GP prescribing of cholesterol drugs (statins), they found huge regional variations and identified £200 million in potential annual savings if cheaper generics were prescribed over brand-names.
- **Shoothill,** which uses the Environment Agency's open flood and river level data to create interactive mapping services to keep the public informed about flood risks. Its FloodAlerts and GaugeMap platforms visualise live data from over 2,400 monitoring stations on easy-to-use maps

Since then, the UK has fallen behind while others have seized the baton and run with it. For example, the EU is creating data spaces for specific high-growth industries, while the US is benefitting from tech leadership that invests heavily in R&D. Unlocking more data from the treasure troves that exist within our state systems will make these isles more attractive for AI development.

With the passage of the Data Use and Access Bill which lays the groundwork for a further opening up of data to unlock smart data schemes, there are signals that the tide is turning once again.

The National Data Library offers the opportunity for the UK to recapture the crown and build on the strong legacy it has in this space.

Challenges and Barriers

Limited and Fragmented Datasets

UK startups find that high-quality data is scarce and siloed. As the Tony Blair Institute for Global Change (TBI) and The Entrepreneurs Network (TEN) noted, Britain's data infrastructure is "built around outdated assumptions" that is "fragmented and unfit for purpose," with public-sector data "locked in silos" that complicate access.[4] In practice, this means many useful datasets either don't exist or are hidden from sight. Founders must often painstakingly comb through dozens of government websites or make time-consuming requests to find basic information, like school attainment records, regional health stats, and local crime figures, leaving little room for creative AI development.

Outdated Formats and Poor Quality

Even when data exists, they are often out of date or in unusable formats. The PAC Committee's *AI in Government* inquiry warned that "out–of–date legacy technology and…poor quality data" in government systems is putting AI adoption at risk.[5] Many datasets are published as PDFs or scanned reports rather than structured files, and agencies rarely invest in cleaning or updating them.

Case Study: Nevermind Global

Startups typically spend weeks wrangling and validating data, time that would be better spent on building models. Nevermind Global, an agentic AI built for transport and logistics, has had some projects forced into the slow lane because of this problem. What should take minutes - identifying EV rapid charger locations - was only possible after spending two weeks cleaning government datasets on the UK mobility data.

This friction has real consequences for startups. While Nevermind already understands the systems they're trying to improve and how to navigate around fragmented access to data, others do not. The barriers aren't capability-related; they're access-related, and they force teams to lose momentum, delay funding cycles, and slow down or shelve innovations that could be game changers. Until public sector data is maintained, regularly updated, and actually clean for use, AI solutions cannot operate at the speed that would most benefit the UK.

[4] https://institute.global/insights/tech-and-digitalisation/governing-in-the-age-of-ai-building-britains-national-data-library

[5] https://publications.parliament.uk/pa/cm5901/cmselect/cmpubacc/356/report.html#:~:text=AI%20relies%20on%20high%20quality,barriers%20are%20persistent%20and%20long%E2%80%93standing

Discoverability and Gaps in the Metadata

Relevant datasets are hard to discover and interpret. For example, a recent study found UK health datasets "rarely include identifiers… or field-level metadata," so their content and suitability are opaque.[6] In online searches, health data findability was no better in 2021 than it was in 2018. In general, the lack of harmonised metadata means startups often don't know what data exists or how to combine it. This "insight-poor" landscape forces innovators to waste time matching up formats or simply rebuilding data from scratch.

Complex Governance and Slow Access

Data-sharing rules vary wildly by sector, creating confusion. Health data, for example, is governed by one set of laws (NHS Act, GDPR, common law duties), scientific data by another (Research Councils/UKRI policies), and general public sector data by yet others (FOI, data protection). This patchwork means obtaining approvals often involves multiple, duplicative processes. One assessment found that even with clear legal paths, "accessing public sector data remains slow and fragmented," with "multiple departmental data-access committees [creating] duplication, inconsistency and unnecessary delays". In short, startups face a bureaucratic maze: each dataset requires its own negotiation, and any mistake can trigger legal issues down the line.

Regulatory and Legal Uncertainty

Finally, the UK's copyright and data laws are in flux, putting the brakes on the UK's ambitions to be an AI powerhouse. At present, UK law only permits text and data mining of copyrighted works for non-commercial research, whereas jurisdictions such as the US, Singapore, Japan and EU have fair use doctrines or have updated their laws to introduce text and data mining exceptions to clarify that data analysis such as AI training is permitted on copyrighted works. The government acknowledges that this "legal uncertainty is undermining investment in, and the adoption of, AI."

Firms have reported to us that a lack of clarity over how they can legally access training data is stunting their ability to grow and innovate - making it more timely and costly. This anecdotal evidence is supported by new polling data of developers, investors, and others working in the UK AI ecosystem from the CCIA.[7] They found that 76% of respondents thought that if the UK chose not to introduce an equivalent protection for text and data mining to those in the EU, US and Japan, it would act as an important signal and the sector would likely reconsider whether the UK is a competitive environment for AI investments.

Until copyright law explicitly allows commercial TDM and clarifies data analysis rights, entrepreneurs are forced to expend scarce resources on lawyers rather than code. This leaves the UK playing on hard mode while other jurisdictions race ahead on easy mode.

[6] https://pmc.ncbi.nlm.nih.gov/articles/PMC8867248/

[7] https://ccianet.org/news/2025/05/uk-ai-ecosystem-poll-shows-the-importance-of-copyright-and-ai-regulation-for-the-governments-objectives-of-promoting-uk-innovation-and-economic-growth/

A 3-Part National Data Library

Part One: Public Sector Data Library and Access

Why?

The UK has historically been viewed as a leader in making public sector data accessible. Startups have benefited from the standardisation of metadata in public sector data, as well as the introduction of open licences to make datasets more easily accessible and reusable.

Startup innovation has moved on significantly since, and founders can now achieve more with public sector datasets than previously thought possible. That is, if they can access the right data.

Startups need consistent access to a variety of publicly-held data sources that are available, harmonised, and findable. Many of the country's startups and scaleups aim to rely on public sector datasets for a variety of uses, including:

- Testing the viability of their products and services
- Tackling real-world problems efficiently and effectively
- Creating updates that work better for the consumer

Case Study: EarlyBird AI

Earlybird builds AI-powered tools for local and combined authorities, employment providers, and government partners to improve frontline public services. Its platform captures rich, structured insights directly from citizens, identifies complex support needs and automatically generates case notes, action plans, and follow-ups to reduce admin and improve outcomes. While Earlybird collects its own primary data, its model works best when connected to structured public datasets such as labour market trends, service directories, and indicators of deprivation. This enables more accurate risk-flagging, personalised support, and faster access to services. However, inconsistent access to timely, interoperable public data across regions limits Earlybird's ability to deliver joined-up support at scale. A national data library could address these gaps and empower responsible innovation across the public sector.

What?

UK startups could unlock significant value if government-held datasets were made more accessible and interoperable. Tax-focused ventures would benefit from richer HMRC VAT feeds and the new Making Tax Digital records. Social-impact founders need a unified API that links police crime references with Home Office statistics to combat injustice and recidivism, and ag-tech firms are eager for DEFRA's farm-subsidy, land-parcel and LandIS soil layers, bolstered by satellite imagery from the Geospatial Commission. More precise weather, solar-irradiance and supplier-level emissions figures (currently only macro) would further accelerate ClimateTech deployments.

On the built-environment side, council-level house-sale surveys, a linkable Land Registry and local-authority rent payment data would allow PropTech and credit risk startups model housing and infrastructure in real time. The NHS and local authorities could catalyse a wave of digital-health innovation by standardising electronic health-record formats and granting wider access to anonymised or pseudonymised datasets. A government-curated, copyright-cleared version of “The Pile” would lower computing costs and legal uncertainty for AI founders while expanding to Commonwealth languages over time. Finally, anonymised electoral rolls, DVLA licence registrations and an open Postcode Address File remain high-priority building blocks for fintech, demographic analysis and logistics startups alike.

How?

However, when making these datasets accessible, it is not enough to just send them out into the digital ether. **We recommend creating a £50 million funding pot housed in DSIT that all departments can bid into to prepare their datasets for release.** Acknowledging that this will vary on a case-by-case basis, public sector datasets will need to fulfil the following three criteria in order to benefit startups:

- Continuously Availability: Are they as easily accessible as, say, the Eurostat or World Bank datasets? Do they not require consent from the Government or specific Civil Service stakeholders to access? Can queries be made to a data service in a way that can handle multiple requests?
- Continuously Harmonisation: Are they updated at timely, regular intervals? Are they standardised in a way that enables them to be combined with other datasets?
- Continuously Findability: Are they able to be found at any hour without assistance? Is their location resilient to cyberattacks? Can they easily be found in a single location? Are there steps that can be taken to enhance indexing or metadata to further improve data discoverability?

Anecdotally, many founders also focused on how a variety of datasets have geographic and historical gaps and are impossible to cross-analyse with other datasets due to their lack of granularity.

All data needs to be standardised for cross-sector applications, calling for increased granularity across the board.

According to the founders, there are two critical departments – HMRC and the Home Office – that need to open up their data in general, as startup founders told us time and time again that they:

- Hold the most valuable data when anonymised
- They are the hardest to access due to a lack of data sharing, even with other public sector institutions

Case Study: Untied

Untied is a specialist fintech streamlining personal tax management and embedding tax services via third-party APIs. It has played a central role in testing the Making Tax Digital for Income Tax programme, showing how automation, embedded finance, and real-time reporting can reduce admin burdens, improve compliance, and support individuals and small businesses.

But like many innovators in the space, *Untied* is constrained by limited access to the granular and anonymised HMRC datasets needed to build smarter, more proactive tools. HMRC holds some of the most valuable and actionable data in government. However, sharing remains limited even with trusted partners, despite the benefits to users and government alike.

Untied's experience shows the potential of meaningful public-private collaboration. If commercial structures enabled HMRC to work more openly with companies like Untied, it would unlock innovation that directly aligns with the government's goals of economic stability, reduced tax gaps, and progress on digital transformation.

Part Two: Opening Up Research and Scientific Data

Why?

Open, publicly funded research data are a public good that startup innovators desperately need. When scientists publish data and code (whether independently or "embedded" in an article) under open licences with standardised metadata, research becomes more transparent, reusable and efficient. Startups can then build on cutting-edge science without reinventing the wheel. Ultimately this means better outcomes for patients who can benefit from the fruits of better research.

For example, Foresight, a generative AI model has the potential to transform patient care, identifying opportunities where early interventions might significantly improve or save lives using national-scale routinely collected, de-identified NHS data, like hospital admissions.[8]

The Comprehensive Spending Review earmarked funding into the life sciences and AI sectors to transform drug discovery and deliver innovation into the NHS frontline. Notably, however, this will have limited success if not matched with greater access to data.

[8] https://www.ucl.ac.uk/news/2025/may/ai-model-trained-de-identified-data-57-million-people

Case Study: Isomorphic Labs

DeepMind's *AlphaFold2*, a landmark AI system for predicting protein structures, was trained on the Protein Data Bank, a publicly funded repository containing over 200,000 biochemical experiments. This breakthrough wouldn't have been possible without open scientific data. It not only accelerated fundamental discovery in biology but also laid the groundwork for DeepMind's next step: *Isomorphic Labs*, a DeepMind spinout focused on reimagining drug discovery through AI.

Isomorphic Labs exemplifies how public data, private R&D, and commercialisation can work together. The spinout retains a strategic link to DeepMind, giving it access to technical support while gradually offloading operational costs: a structure designed to maximise both sustainability and impact. Crucially, it shows how open science fuels commercial innovation, creating space for new companies to emerge from foundational research.

This model reflects what's possible when open, standardised research data is treated as a national asset: it becomes the launchpad for globally competitive startups that blend scientific insight with AI capability.

Startups rely on open research data for many purposes, for example:

- Open genomic, clinical and biochemical datasets enable biotech and pharma startups to accelerate drug discovery and medical innovation.
- Access to environmental, weather and emissions data lets climate-tech companies model impacts and optimise energy systems.
- AI startups need vast corpora (e.g. text, images, scientific data) to train highly-specific models. More openly licensed data would allow innovators to reuse taxpayer funded research outputs freely.

What?

In practice, we must require FAIR data practices that are rich, machine-readable metadata and interoperable formats as well as the application of open licences (e.g. CC-BY or the UK Open Government Licence) to all datasets and publications. Startups have repeatedly told us that traditional academic publishing models (with long embargoes) slow down innovation. Introducing a secondary publishing right would address this as it would allow researchers to deposit publicly funded articles in repositories immediately upon publication. Combined with open-data mandates, this would create an "OA by default" regime where any research output can be found and reused by startups without legal barriers.

In addition, the introduction of a text and data mining exception would also provide clarity demonstrating that legally accessed articles can be analysed and used for AI, accelerating innovation and scientific research.

Ultimately, the goal is a unified system of open science. By making public research datasets findable, accessible and interoperable – and available for reuse – we empower startups across life sciences, ClimateTech, and AI. Crucially, these changes would mean that biotech firms are harnessing NHS and genomic data securely, climate innovators are using high-granularity environmental records and AI models are trained on truly representative UK data. To achieve this, we must lock in current gains: UKRI and other funders should continue the move to open licensing and data-sharing plans as a funding condition. The KR21 position paper shows Europe-wide momentum for zero-embargo secondary publishing, reinforcing the case for domestic reform.[9]

How?

To translate these principles into action, policymakers should adopt clear mandates and incentives. In particular, funding rules and legal reforms must enforce open data. For example, all UK research grants should require detailed Data Management Plans and make publication of underlying data a condition of funding. Final grant instalments could be withheld until datasets are publicly deposited, ensuring compliance as standard practice.

Similarly, existing copyright and licensing regimes should be clarified. We recommend introducing a statutory Text & Data Mining exception for all AI and commercial research (not just non-commercial use) to enable analysis and AI use, and securing a UK-wide secondary publishing right to eliminate publication delays. By combining these with open license mandates, every public sector dataset and paper would be reusable by default.

- Require all publicly funded research outputs to be deposited in trusted, accessible repositories (with long-term funding). Data must be FAIR: richly described, machine-readable and interoperable as well as being released under CC-BY or OGL licences.
- Give researchers and commercial entities permanent rights to mine and redistribute research data. Match the UK's text and data mining approach to other leading jurisdictions and foster a zero-embargo secondary publishing right for all publicly funded work.
- Make open-data delivery a grant condition. Funders (UKRI, NIHR, etc.) should verify that datasets are released before awarding final payments, and integrate data-access metrics into evaluation (e.g. the Research Excellence Framework).
- Allocate dedicated funding (on the order of approx. £25m) to upgrade research data tools and repositories. This could be a DSIT-administered challenge fund to help universities and labs clean, curate and host their data to modern standards.
- Organise challenge prizes or hackathons (e.g. an "AI Reuse" competition) to spotlight innovative startups that leverage open research data. Celebrating success stories will demonstrate the economic impact of open science and inspire others.

These steps, together with ongoing public-sector data efforts, will ensure UK startups in life sciences, ClimateTech, AI and beyond can fully exploit the wealth of publicly funded research. Startup Coalition believes that opening up science data in this way is a key enabler of a thriving innovation economy.

[9] https://www.knowledgerights21.org/wp-content/uploads/KR21-Secondary-Publishing-Rights-Position-Paper-September-2024.pdf

Part Three: Horizon Scanning and Partnerships

Why?

Emerging startups thrive on data, but not all relevant datasets exist today. We need to anticipate future data demand or the UK risks missing the next big breakthroughs. Founders building AI models, ClimateTech, and health solutions rely on new data, not just what's accessible right now.

To best help startups aiming to deliver public value, the NDL must shift from a data-first mindset to a mission-led and challenge-driven approach. Aggregating data that could be useful in these scenarios still limits its use. Without a unifying vision or problem to solve, these data initiatives can become fragmented and underutilised, causing costly policy lag.

In fact, many have already described what founders fear: that focusing on datasets alone could lead to the NDL lacking focus and devolving into a data repository that cannot be successfully harnessed for good. To best help mission-driven founders, data efforts should not be the end in themselves, but the means to help solve tangible societal problems, regardless of whether the data sets currently exist.

By identifying public missions and challenges upfront, the NDL can directly align its work to support the Government's priorities. This will also foster cross-departmental and cross-sector collaboration, which would provide much-needed standardisation when it comes to data as a byproduct. As the TBI and TEN report described, the NDL could be a real deliverer of the Government's plan for change. And a mission-led NDL would clarify the NDL's purpose and secure its political support for long-term projects.

What?

In practice, a challenge-down model would start with specific public policy problems or missions and use the NDL to work towards the data and tech needed to solve them, in conjunction with startups and the tech sector itself. That would make data the byproduct of solving concrete and knotty problems. Ways to this would include challenge prizes - similar to the Smart Data ones organised by the Department of Business and Trade and others which allow startups to test and prototype solutions that cut across traditional sector boundaries. While the Government and its partners define the priority challenges, they work with startups to solve them and create public-private foundations to generate more interoperable data across sectors.

A challenge-driven mode would need to focus on the following elements to succeed:

- A partnerships-first data strategy that brings startups, researchers, and Government together around Labour's missions or increasing problems faced by the UK public
- Plug-and-play standards for data that ensure that the data and outputs used in one challenge can be reused later on, essentially creating the data infrastructure needed moving forward
- Mission boards and challenge teams to set direction, track delivery, and ensure the data is in the details
- Ways to scale these open calls and challenge prizes into programmes that deliver data-driven innovation

Case Study: Our Future Health

Our Future Health is a mission-led initiative to improve disease prevention and treatment by gathering health data from up to five million UK volunteers. By setting the clear goal of accelerating medical breakthroughs this public-private partnership is actively collecting an unprecedented scale of health information (from genomic to lifestyle data).

In doing so, it demonstrates how focusing on a societal challenge can mobilise data generation and sharing. The rich dataset emerging from *Our Future Health* will fuel AI-driven healthcare research for years to come, showcasing the power of a challenge-driven approach.

How?

- Publish a challenge-based roadmap that sets out a clear, public plan that links national missions (e.g. Net Zero, health, economic resilience) to real-world data challenges. Include short-term use cases and long-term goals. Make it obvious where the value is and what data is needed to unlock it.
- Create cross-cutting boards to own each mission. Give them the remit, people and authority to set the agenda, coordinate delivery, and unblock data access across departments and sectors. These should be drivers, not observers.
- Run regular challenge prizes aligned to mission goals, backed by sandboxes with real or synthetic data. Surface ideas from startups, researchers, and civil society and fund what works. Crucially, make it easy to test, iterate, and scale.
- Mandate interoperability from day one by requiring shared standards, open APIs, and reusable data models across sectors. Convene working groups to agree on common formats, and link compliance to access, funding or procurement.
- Bake data-sharing, standards, and collaboration into government contracts, grants, and R&D funding. Support organisations to do it well by working with templates, playbooks, and platforms. Build a system where data gets shared by default when it serves the public good.

Conclusion

The successful implementation of a National Data Library hinges on actionable workstreams that enforce greater and more consistent access to vast, context-specific, and relevant data. By making data more accessible and reusable, UK startups can advance innovation and drive progress across sectors.

This could be achieved by:

1. Opening Up Government Held Datasets

- Creating a £50 million funding pot housed in DSIT: This fund would allow departments to prepare their datasets for release.
- Ensuring continuous availability, harmonisation, and findability of datasets: This means datasets should be easily accessible, updated regularly, standardised for combination with other datasets, and discoverable in a single, resilient location.
- Increasing granularity of data: More detailed data is needed for cross-sector applications and high-quality analytical models.

2. Opening Up Research and Scientific Data

- Requiring FAIR data practices and open licenses for all publicly funded research outputs: Data should be richly described, machine-readable, interoperable, and released under open licenses (e.g., CC-BY or UK Open Government Licence).
- Introducing a statutory Text & Data Mining exception for all AI and commercial research: This would clarify legal access to copyrighted works for data analysis.
- Securing a UK-wide secondary publishing right: This would eliminate publication delays by allowing researchers to deposit publicly funded articles in repositories immediately upon publication.
- Making open-data delivery a grant condition: Funders should verify dataset release before awarding final payments and integrate data-access metrics into evaluation.
- Allocating dedicated funding (~£25m) to upgrade research data repositories and tools: This challenge fund would help universities and labs clean, curate, and host their data to modern standards.
- Organising challenge prizes or hackathons: These would spotlight innovative startups leveraging open research data.

3. Horizon Scanning and Partnerships

- Publishing a challenge-based roadmap: This roadmap should link national missions to real-world data challenges, outlining short-term use cases and long-term goals.
- Creating cross-cutting boards for each mission: These boards would set the agenda, coordinate delivery, and unblock data access.
- Running regular challenge prizes aligned to mission goals: These should be backed by sandboxes with real or synthetic data to encourage testing, iteration, and scaling of solutions.
- Mandating interoperability from day one: This includes requiring shared standards, open APIs, and reusable data models across sectors.
- Baking data-sharing, standards, and collaboration into government contracts, grants, and R&D funding: This would make data sharing a default practice when it serves the public good.

STARTUP
C✱ALITION