Prototyping an AI-ready National Data Library

Data-centric AI

DAI

ODI Research
ADVANCING TRUST IN DATA

March 2026

Contents

Introduction **2**

Executive summary 4

Related work and the foundations of the NDL **6**

Designing the prototype **9**

Composition 9

Processing 11

Accessibility 12

How to use the NDL-lite **13**

Insights **17**

It's not difficult to get started 17

Public sector data repositories will hold the NDL back if they are not AI-ready 18

AI agents find government data hard to use, often choosing to look elsewhere 19

Conclusions and recommendations **21**

Appendix: Technical specifications **23**

About

This report was researched and produced by the Open Data Institute and published in March 2026. Its lead author was Neil Majithia, with support from Huseyin Kir, Elena Simperl and Emma Thwaites. If you want to share feedback by email or would like to get in touch, contact the research team at research@theodi.org

Licence:

Introduction

The UK's AI Opportunities Action Plan,[1] released in January 2025, sets out the government's ambition not only to support national AI innovation but also to capture the benefits of that innovation by using AI to empower public services, research, and overall growth. The National Data Library (NDL), a key facet of the action plan, can fulfil both of these ambitions simultaneously.

Firstly, the proposed NDL acknowledges just how important access to data is for innovation: there is no AI without data.[2] If the NDL contains the swathes of public sector data collected across the UK (including economic, scientific and environmental datasets held by public bodies), the government can counteract an increasingly expensive and inaccessible AI data ecosystem[3] that 'prices out' UK-based startups and academic institutions in favour of large, entrenched entities from the US and China. In keeping with the UK's historically renowned approach to open data,[4] the NDL presents a way to support homegrown talent with the datasets at its disposal, fostering a competitive edge in the world of AI.

Secondly, the NDL puts the UK in a position to capitalise on AI's benefits. Sophisticated AI agents like Claude Code[5] and Gemini-CLI[6] have revolutionised the world of programming by enabling users to write entire codebases in seconds.[7] Research, modelling, forecasting and product development promise to be easier and faster. With the NDL making public sector datasets centrally accessible, startups, academics and civil servants can use AI agents to help detect hidden patterns across the UK's data landscape, building intelligence and products to take public services and policymaking into a new era.

However, while its potential is clear and government backing is in place, the NDL itself has publicly remained mostly abstract, an idea rather than an implementation. The January 2026 progress update[8] does outline the

---

[1] Department for Science, Innovation and Technology (2025), 'AI Opportunities Action Plan'
[2] Snaith, B. (2023), 'What do we mean by 'without data, there is no AI'?'
[3] Paul, K. & Tong, A. (2024), 'Inside Big Tech's underground race to buy AI training data'
[4] Carrara, W., Fischer, S. & van Steenbergen, E. (2016, revised 2020), 'Open Data Maturity in Europe 2015 | Insights into the European state of play'
[5] Anthropic (2025), 'Claude 3.7 Sonnet and Claude Code'
[6] Mullen, T. & Salva, R.J. (2025), 'Google announces Gemini CLI: your open-source AI agent'
[7] Wang H. et al. (2025), 'AI Agentic Programming: A Survey of Techniques, Challenges, and Opportunities'
[8] Department for Science, Innovation and Technology (2026), 'National Data Library: progress update, January 2026'

evidence gathering activities that the Department for Science, Innovation and Technology (DSIT) has conducted behind the scenes, and provides five 'kickstarter projects' for development to focus on. But it has been almost a year since the NDL was proposed and few concrete steps have been taken towards its actual construction; we run the risk of losing momentum, both with the NDL and with the UK's role in the global AI ecosystem.

The Open Data Institute (ODI) has already put forward a conceptual architecture that we believe the NDL should emulate,[9] focused on how it should be, at its core, 'AI-ready'. We have further defined what 'AI-readiness' looks like in our Framework for AI-ready data,[10] which sets out several criteria that datasets, metadata, and surrounding infrastructure should meet to ensure an overarching data product like the NDL is considered (socio)technically optimal for AI. DSIT recently cited our framework and built on it for its own 'Guidelines and best practices for making government datasets ready for AI'.[11]

In this work, we now put our words into action by taking the aforementioned research and, with the ODI's institutional expertise and experience, designing and implementing **NDL-lite**, an AI-ready prototype of the NDL that represents the first step towards its practical implementation. In doing so, we learnt about the challenges the government will face in data collection, curation and processing across multiple sources, and built three key insights that should be taken into account when the NDL initiative reaches its development stage.

The NDL-lite that we are releasing openly contains 38GB of data from six UK public sector repositories, encompassing both structured and unstructured datasets from the fields of economics, law, policy, public administration, science and the environment. With suitable metadata and a comprehensive architecture that facilitates search, download and usage of datasets in AI workflows, the prototype already enables any researcher, armed with an AI agent of their choice, to access and analyse UK public sector datasets or use them to train algorithms and models. This report provides an example in which a user works with an AI agent to analyse statistics on electric vehicle charging infrastructure in rural areas against transcripts of parliamentary debate, building evidence to identify an entrepreneurial opportunity that meets both political objectives and rural communities' needs.

---

[9] Meroño-Peñuela, A. et al. (2025), 'How an AI-ready National Data Library would help UK science'
[10] Majithia, N., Carey-Wilson, T., and Simperl, E. (2025), 'A framework for AI-ready data'
[11] Government Digital Service & Department for Science, Innovation and Technology (2026), 'Guidelines and best practices for making government datasets ready for AI'

We built our prototype with reusable scripts and data pipelines that provide examples of the kind of methods that the NDL initiative should employ. To that end, we provide all code and documentation in the open-access GitHub repository here.

This report contains an overview of our work on this prototype. It includes our considerations of its design and potential use and our comments on the practical experience of its construction. These elements combine into a rich set of constructive conclusions and recommendations that can be adopted by the teams looking to move forward on both the NDL initiative and the UK's national data infrastructure.

Executive summary

We designed the NDL-lite by referencing swathes of research and policy reports about the NDL initiative and what different stakeholders believe it should entail.

In total, the prototype contains 38 gigabytes of data from six public sector data sources (data.gov.uk, gov.uk, environment.gov.uk, hansard.parliament.uk, legislation.gov.uk and ons.gov.uk), each permitting data reuse with the Open Government Licence. We aggregated more than 100,000 files from these sources, processing them to clean and standardise them, before ensuring accessibility with a sophisticated 'access layer' that enables interactions with agentic AI.

We demonstrate some examples of how the NDL-lite can be used, then present three insights we found over the course of development:

- It's not difficult to get started
  - To build a successful prototype, our team worked with limited resources and methodologies, many of them based on open-source infrastructure. We argue that this demonstrates the ease of building something tangible and taking a first step towards converting the NDL from a policy ideal to a practical reality.
- Public sector data repositories will hold the NDL back if they are not AI-ready
  - Something that holds the utility of the NDL-lite back is that many datasets, especially those aggregated from data.gov.uk, are simply difficult to use. With misleading titles and non-existent metadata, these datasets cannot support any meaningful analysis. We recommend that any NDL work addresses improving the state of public sector datasets before they are made accessible through the NDL.
- AI agents find government data hard to use, often choosing to look elsewhere

- Our experiments found that the cutting-edge AI agents revolutionising technical research and development are prone to looking for data outside of government sources if they find government data hard to use. For example, they prefer to search the internet for national crime statistics rather than use any of the government datasets incorporated in the NDL-lite, which they find difficult to use. For these agents to be incentivised to use only trusted, authoritative government data sources, we believe it is the government's responsibility to ensure those data sources are robustly AI-ready.

We conclude by noting that though the NDL-lite is a work-in-progress, it is a tangible, affordable and fast first step that represents a movement away from concepts and towards realities. By publishing both the prototype and the code underlying its creation publicly, we hope to provide a focus point for redesign and iteration towards a fully functional NDL.

Related work and the foundations of the NDL

The concept of a data library has existed for decades. Centralised repositories for datasets enable their sharing across institutions and organisations, empowering (inter)national collaboration for groundbreaking research.

Take, for example, the Protein Data Bank (PDB).[12] For more than half a century, the PDB has enabled bioinformaticians from across the world to share, visualise and peer-review almost every discovered protein sequence. Armed with this information, laboratories have been able to design, synthesise and then mass-produce proteins that revolutionise biomedical engineering.[13] The PDB was the data source used to train AlphaFold,[14] Google DeepMind's series of Nobel-prize winning[15] AI models that have enhanced cancer drug discovery,[16] provided faster emergency medicine in pandemics,[17] and supported the development of malaria vaccines.[18]

Alongside other data repositories such as the UN Humanitarian Data Exchange (HDX),[19] the PDB exemplifies the obvious potential benefits an NDL could bring to AI. However, the PDB is focused purely on the single domain of bioinformatics; the NDL would go even further given its proposed cross-disciplinary nature and the full breadth of data that the UK has to offer.

Importantly, the UK already has strong public data foundations to build upon. The Office for National Statistics (ONS),[20] the Department for Environment, Food and Rural Affairs (Defra),[21] and the National Health Service (NHS)[22] are just three of the public institutions that publish open data, and data.gov.uk, an openly accessible central resource for public sector datasets, was once a blueprint for how international governments could build their own data services. All public data on data.gov.uk is published with the bespoke Open

[12] Research Collaboratory for Structural Bioinformatics Protein Data Bank (n.d.), 'RCSB PDB: Homepage'
[13] Open Data Institute (2025), 'The Open Data Use Case Observatory | Open Data Accelerates Nobel Prize-Winning AI in Molecular Research'
[14] Google Deepmind (n.d.), 'AlphaFold'
[15] ChemH (2024), 'Beyond AlphaFold 3: Navigating Future Challenges in Protein Structure Prediction'
[16] Ren, F. et al. (2023), 'AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor'
[17] Higgins, M.K. (2021), 'Can We AlphaFold Our Way Out of the Next Pandemic?'
[18] Google DeepMind (2022), 'Stopping malaria in its tracks'
[19] UN OCHA Humanitarian Data Exchange (n.d.), 'Humanitarian Data Exchange'
[20] Office for National Statistics (n.d.), 'Published data related to economy'
[21] Department for Environment, Food and Rural Affairs (n.d.), 'Data services platform'
[22] NHS Business Services Authority (n.d.), 'Open Data Portal'

Government Licence and, consequently, the UK government has become renowned as a leader in the open data space.[23]

However, given the insufficiencies of government data platforms like data.gov.uk,[24] there is a need to shift from current, outdated public data infrastructure to an innovative NDL.

The plan for this shift was first officially laid out in the AI Opportunities Action Plan,[25] in which the Secretary of State for Science, Innovation and Technology introduces the NDL concept and proposes an initial iteration that would cover at least five 'high-impact' public datasets 'in strategically significant areas, building on existing UK strengths'. The plan omits further technical detail in favour of proposing the governance and policy surrounding the NDL, including incentives, provision of compute resources, and strategy at an international level.

The action plan's recommendation for an NDL has been adopted in official policy. The Blueprint for Modern Digital Government[26] lists the library as the third item in its 'Six-point plan for public sector digital reform', and in the 2025 Spending Review,[27] the government committed funding for DSIT to construct it.

To collect suggestions for the actual architecture to underpin these plans, Wellcome held a technical white paper challenge in early 2025 for UK-based institutions to contribute their expertise towards the design of the NDL.[28] Submissions[29] from the Bennett Institute for Applied Data Science, The Francis Crick Institute, DARE UK, Icebreaker One and the ODI each put forward varied technical architectures that the government should consider.

A number of existing data repositories could serve as architectural inspiration. The aforementioned Protein Data Bank's infrastructure and features have crystallised more than 50 years of maturity into a data repository that truly serves innovation. Similarly, Wikibase,[30] the underlying technology behind Wikipedia and Wikidata, provides means to store masses of structured data in a way that facilitates semantic interoperability and, now, access with AI.[31]

---

[23] Open Data Barometer (2017), 'Open Data Barometer'
[24] Majithia, N. et al. (2024), 'The UK government as a data provider for AI'
[25] Department for Science, Innovation and Technology (2025), 'AI Opportunities Action Plan'
[26] Department for Science, Innovation and Technology and Government Digital Service (2025), 'A blueprint for modern digital government'
[27] HM Treasury (2025), 'Spending Review 2025'
[28] Wellcome (2024), 'UK National Data Library: Technical White Paper Challenge'
[29] Wellcome (2025), 'UK National Data Library: Technical White Paper Challenge – Submissions'
[30] Wikibase (n.d.), 'Wikibase'
[31] Wikidata (2026), 'Wikidata:Embedding Project'

More recently released datasets that were explicitly made for the purposes of AI training also provide precedent. The Common Pile,[32] an eight-terabyte dataset released by EleutherAI in Summer 2025, contains high quality, standardised text data from 30 diverse open data sources, encompassing nine topical domains (such as code, law and academia). While this dataset is much smaller than those typically used to train LLMs, it is extremely popular, having proven to deliver state-of-the-art performance for AI models trained with it while – importantly – being completely openly accessible and responsibly collected. CommonCorpus,[33] released by Pleias, also provides a cross-discipline real-world knowledge base consisting entirely of open data, enabling the training of small, high-quality AI models that are compliant, transparent, and provably safe for use in industry or government.

In the UK, the London Datastore[34] is perhaps the closest approximation of the starting point for the NDL initiative. It was developed with a similar aim: to centralise, publish, and encourage usage of datasets involved with the governance of London. The Datastore covers 18 topic areas, with thousands of datasets accessible via an online portal or API. However, there is no standardisation or interlinkage between these datasets, nor are they published in AI-ready formats (most are only downloadable in .xlsx, a proprietary format that is inefficient for innovators to use). A national expansion of the datastore would therefore not accomplish the aims for the NDL initiative, nor would it resemble any suggestions from the white paper challenge.

The NDL requires a prototype built on the evidence presented in this section, a first step that the government can start with, build on, iterate and scale into an overarching product that suits the proposal laid out in the AI Opportunities Action Plan – a national resource that puts the UK at the forefront of the global AI ecosystem.

[32] Kandpal, N. et al. (2025), 'The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text'
[33] Langlais, P.-C. et al. (2025), 'Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training'
[34] London Assembly (n.d.), 'London Datastore'

Designing the prototype

Informed by the swathes of related work presented in the previous section, we set out to build a collection of UK public sector data that has been aggregated, structured, processed and published in a way that supports AI workflows. We therefore had three core principles:

- **Composition:** aggregation of UK public sector datasets alongside attached metadata.
- **Processing:** sufficient standardisation and cleaning of aggregated datasets and metadata.
- **Accessibility:** structure and publication in a way that is suitable for all manners of AI interactions,[35] including agentic AI.

Composition

From a longlist of 30 prospective public data sources, for this initial release we decided to prioritise six that we deemed the highest-value and most all-encompassing for aggregation in our prototype, in accordance with suggestions in the Action Plan.[36] The data sources, their types, and their representation in the prototype are detailed in Table 1.

Data source	Type	Volume	Comments
data.gov.uk	Structured Data and Text	34.5 GB (37,172 files)	We collected the 300 most recent publications across the 13 categories of data on data.gov.uk
gov.uk	Text	2.44 GB (67,718 files)	Using the gov.uk API, we accessed all core gov.uk website information as well as all publications from 2025.
environment.gov.uk	Structured Data	1 GB (933 files)	The 300 most recent publications combined to give 933 datasets.

[35] Hardinges, J. & Simperl, E. (2024), 'A data for AI taxonomy'
[36] Department for Science, Innovation and Technology (2025), 'AI Opportunities Action Plan'

hansard.parliament.uk	Text	146 MB (826 files)	We aggregated all content, including debate transcripts and written Q&As, published in 2025.
legislation.gov.uk	Text	100 MB (1,925 files)	We aggregated all content published, updated, or otherwise modified in 2025.
ons.gov.uk	Structured Data	130 MB (22,151 files)	Restricting to time series data only, we aggregated the 15 most recent datasets across 10 topic areas. The ONS has an API usage limit that severely restricts the amount of data we could ingest from it. This means that some core datasets that would be expected in the NDL are not present in our prototype.

Table 1

Our rationale for choosing these data sources was that, once centralised and pre-processed, they together can be analysed to discover patterns that would enable innovation in public and private sectors or lead to better-evidenced policymaking. Each can enrich another: for example, analysing economic statistics from the ONS against environmental data from Defra, then contextualising findings with text from legislation or parliamentary debate, can produce greater, deeper insights than using any of these data sources by themselves.

Processing

The sheer variety in formats and structures of publishing formats among the aggregated public sector data sources required a sophisticated pipeline for data refinement and processing. We used the criteria of our Framework for AI-ready data as guiding principles in the design of this pipeline, which includes:

- **Cleaning:** all structured data is checked for duplicates and null values. Dates and times are converted to ISO 8601 standard format. Any Personal Identifiable Information (PII), such as phone numbers or email addresses, are redacted.
- **Structuring:** unstructured data formats like PDFs are converted to text using Optical Character Recognition (OCR) tools. Text in HTML is also extracted into a plain text format. Structured datasets, published as CSVs, XLSXs, JSONs and more, are standardised into the Apache Parquet format to optimise their performance for analytical workloads.
- **Enrichment:** each file is augmented with metadata, including automatically detected language tags and EU Data Theme[37] vocabulary tags, to ensure contextual clarity and semantic consistency across the corpus. We preserved the small amount of metadata published alongside the datasets we aggregated.

The final component of our processing was the generation of vector embeddings. This is a way of representing text – such as the title and description of datasets – as a list of numbers that mathematically communicate its semantic 'meaning'. AI algorithms can use these lists of numbers to search for datasets within the NDL prototype in a more sophisticated manner than traditional keyword search, enhancing the utility, ease of use, and machine-accessibility of our work. These embeddings are also the foundations upon which retrieval-augmented generation (RAG) AI chatbots, like GOV.UK Chat,[38] can be built to ensure answers to user queries are grounded in real government data.

[37] E-Government Standards (n.d.), 'Theme vocabulary'

[38] Government Digital Service (2026), 'govuk-chat: A web application that provides a LLM powered chat experience based on GOV.UK content.'

Accessibility

With the two previous steps completed, we designed and organised the publication of our prototype in a means that is accessible to humans and AI alike. There are three distinct published datasets:

- **NDL-lite Corpus:** the collection of all aggregated and processed text data, served in chunks of 800 characters (link).
- **NDL-lite Structured Data:** the collection of all aggregated and processed structured data, served as Parquet files (link).
- **NDL-lite RAG Index:** the dedicated index of embeddings for all files in both datasets, acting as a database for semantic search (link).

However, were we to stop at simply publishing these datasets, we would not be truly supporting innovation. We ensured we built an 'access layer' to easily use the datasets we aggregated. This has two components:

- **Data Application Programming Interface (API):** we expose two API endpoints that enable a user to easily search for data in the NDL-lite and download it to their working environment. API-based access is a core component of AI-readiness, enabling both human analysts and algorithmic workflows to easily find, access and use data (link). The Croissant records attached to datasets also facilitate their import to common machine learning workflows like Tensorflow and Pytorch.
- **Model Context Protocol (MCP):** in accordance with recent developments in the rapidly evolving world of AI, we set up an MCP implementation. With supporting Python infrastructure, the implementation allows modern-day AI agents (like Claude Code or Gemini-CLI) to programmatically query the NDL-lite. The MCP server exposes tools that enable agents to search for structured and unstructured data and download it entirely within their native chat interfaces. In practice, this means a human analyst hardly has to write any code themselves; they can simply ask their AI agent to find datasets and analyse them automatically (link).

How to use the NDL-lite

By following the documentation in the access layer above, a user can work with our NDL prototype to search across public sector datasets and conduct cross-cutting research, whether by themselves or with the help of an AI agent.

Take, for example, a startup founder looking to determine market need around electric vehicle chargers. They ask an AI agent:

```
USER> How is the rollout of electric vehicle chargers actually
going? Search for statistics on EV charging infrastructure. I want
to see a growth trend of how many devices have been installed in
recent years. Additionally, tell me what policymakers in debates
are saying about terms like 'Net Zero' or 'EV charging' to see if
MPs are complaining about specific issues, like a lack of chargers
in rural areas. Contrast the official stats with the political
reality.
```

The agent spends 30 seconds thinking, then searches the NDL-lite to find statistics about electric vehicles. It finds some in a publication on gov.uk, having determined that they were more recent than those on data.gov.uk, and writes them to a file to then analyse. Afterward, it pulls qualitative Hansard records from the NDL-lite that feature parliamentary debates on the topic.

Finally, the agent synthesises the two data sources it finds to write a report to the user:

```
AGENT> 1. Quantitative Analysis: Growth Trend (2025)
The data indicates a steady growth in the number of public
charging devices throughout 2025.
   * January 1, 2025: 73,334 public charging devices (including
14,448 rapid chargers).
   * February 2025: >74,000 public charging devices.
   * November 1, 2025: >86,700 public charging devices.
Growth: The network grew by approximately 13,366 devices (an ~18%
increase) between January and November
  2025. The government reported a 22% year-on-year increase as of
November.
```

```
  2. Qualitative Analysis: The Political Reality
Despite the growth in raw numbers, parliamentary debates reveal
significant concerns regarding the distribution and reliability of
the infrastructure.
  Key Themes in Debates:
   * Rural vs. Urban Divide: MPs consistently questioned the
"adequacy" of infrastructure in rural and remote areas. The term
"patchy provision" was explicitly used to describe the user
experience outside major hubs.
   * The "Driveway Gap": A major concern is equity for drivers
without off-street parking. MPs pressed the government on
"cross-pavement solutions" and support for terraced housing
residents who cannot charge at home.
   * Grid & Cost: High electricity costs and "inadequate power
capacity" for rapid charging on strategic roads were cited as
barriers to uptake, with some constituents seeing their bills
triple.
Government Response:
  Ministers (e.g., Michael Shanks, Keir Mather) defended the
record by citing the £381 million Local EV Infrastructure (LEVI)
Fund, which aims to deliver 100,000 new local chargepoints. They
frame the issue as a "rollout at pace" that will match demand,
though they acknowledge it is a devolved matter in Scotland.

  3. Conclusion: Stats vs. Reality
 Official Stats: Show a robust, double-digit growth curve and
significant investment (£381m LEVI fund).
  Political Reality: Highlights that coverage matters more than
count. The "total number" statistic masks the reality of
"transport deserts" in rural areas and the practical difficulties
for urban residents without driveways. While the government points
to the future (100,000 planned chargers), MPs are complaining
about the current lack of visible infrastructure in their specific
constituencies.
```

Happy with this report, entirely built upon public sector data in the NDL-lite, the startup founder builds a business plan, working with the agent to refine their idea for a new, better EV charging infrastructure that is purpose-built for the UK.

In another example, a researcher is studying economic productivity in the UK. Seeking to analyse how it interacts with socioeconomics, they start up the AI agent of their choice, hook it into the NDL-lite MCP server, and ask it:

```
USER> I'm looking for data on national productivity. Find a
dataset, and visualise its contents for me.
```

The AI agent receives this request, checks out the MCP server, and immediately gets to work. Using the RAG embeddings we published, it uses our API tooling to search for datasets in the NDL-lite that fit the user's needs. It identifies one: 'Productivity jobs', a dataset release from the ONS on ons.gov.uk. Importantly, since the dataset itself was not published with any metadata or descriptive context (as can be seen on its webpage), the agent first verifies that this dataset is indeed about national productivity as measured by the ONS and seasonally adjusted.

The agent, noticing that the dataset is provided as a Parquet file by the MCP server, then automatically writes code to read the dataset and identify what it contains. It then turns its attention to creating a visualisation, as shown in Figure 1 below:

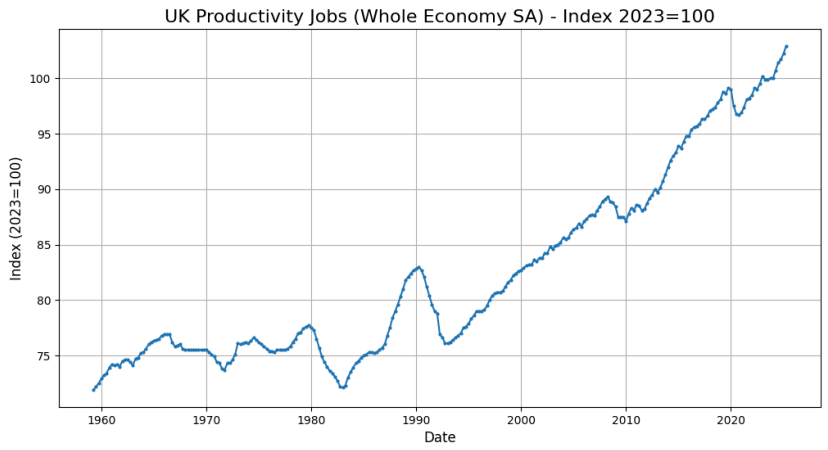

Figure 1: the graph of national productivity drawn by the AI agent.

At this point, the researcher adds another prompt, seeking to confirm an initial hypothesis.

```
USER> Thanks. I remember from my Economics degree that, in an
ideal world, productivity is closely interlinked with (real)
wages, and when one goes up, the other goes up too. With this
dataset and others in the NDL, please visualise whether this
relationship holds, including providing statistics and whether
there has been a "structural break" in the way the measures are
interlinked.
```

In response, the agent quickly searches the NDL-lite, finding two more datasets: one that provides ONS measurements of nominal wages, and another that provides a 'deflator' that the ONS uses to turn nominal wages into real ones. It combines the two, quickly runs correlation analysis, then outputs another visualisation while also providing an interpretation of its analysis.

```
AGENT> The data supports the hypothesis of a "structural break"
around 2008. While productivity and wages moved in lockstep before
the crisis (correlation 0.97), the relationship decoupled slightly
afterwards   (correlation dropped to 0.85), often referred to as
the "productivity puzzle" or wage stagnation relative to
productivity.
```

The user seeks to use more than just ONS data.

```
USER> I'm interested in the hypothesis that this wage stagnation
had effects on the social ecosystem of the UK. Find data on
national crime statistics, or health statistics, or education, to
identify any relationships between them and the delta between
productivity and real wages.
```

Having been given free rein, the agent begins to search for datasets in the NDL-lite. It starts by looking for crime data, finding the 'Crime' and 'Crime Statistics' datasets from data.gov.uk, although it disregards both – the former because it is restricted to the area of Calderdale, and the latter because it contains only data from the years 2011-2018, with many subsets only accessible as archived HTML pages. The agent then searches for pupil absence data, finding datasets but leaving them unused because they only concern the years 2011-14. It seeks health data, but it finds only the hundreds of recently-published audit records and expenditure reporting on data.gov.uk.

In fact, the agent tries for five more minutes before concluding that there are no datasets concerning non-economic outcomes that suit the user's needs among the 60,000 files of structured data in the NDL-lite. After listing some findings from unstructured data sources instead, it reverts to what it knows it can work with – data from the ONS – to compare the wage-productivity gap against labour statistics. When prodded further, the agent makes the decision not to consult the NDL-lite and instead looks for crime statistics on the web.

In this example, the utility of the NDL-lite is held back by the data quality within it. This acts as a cautionary tale that we explore in the following section.

Insights

It’s not difficult to get started

While our work is only a demonstration of a prospective piece of national technological infrastructure, the NDL-lite contains almost 40 gigabytes of data from public sector repositories, covering many subject areas and enabling access to them with cutting-edge AI.

Building this prototype was not difficult. Four people at the ODI, given only public information about the NDL and the ODI’s previous research on the subject, planned, built and iterated the architecture behind the NDL-lite within four months (with only two months of key development work).

The architecture has four components: (1) a robust data ingestion and processing pipeline; (2) centralised storage; (3) a vectorised search index; and (4) the access layer. The technical frameworks underpinning each of these four components, such as Dagster, FastAPI and MCP, are ubiquitous across industry, easy to implement, and extremely low-cost.

Furthermore, most of the tooling we used, including the generator for vector embeddings, was open-source. The open-source AI ecosystem should not be overlooked; it is a nexus of global innovation in terms of both technological advancement and responsible AI data practices.[39]

With this in mind, it is difficult to justify inaction. Building a prototype is a small task and represents the start of an iterative process of the NDL’s development; meanwhile, if we continue to procrastinate before taking such an incremental first step, the NDL will never get off the ground.[40]

A prototype helps to focus discussion. Rather than speaking about blueprints and plans, key stakeholders in the NDL initiative will have something tangible to review, praise or critique if a prototype is put in front of them. Releasing it publicly offers an opportunity to crowdsource possible improvements and the next steps towards truly meeting the needs of innovators across the UK.

---

[39] Such as the Common Corpus or the Common Pile.

[40] We acknowledge that there may be many political and financial factors underlying the seeming halting momentum of the NDL. From a technological perspective, though, there is little standing in the way of progress.

Public sector data repositories will hold the NDL back if they are not AI-ready

Before collecting the datasets that would eventually be a part of the NDL-lite, we sought to assess their sources. We explicitly focused on data.gov.uk, which was once a crown jewel of the UK's open data strategy and comprises a large portion of our prototype.

Using our Framework for AI-ready data, we found that data.gov.uk, as a data product, hardly fits our requirements for supporting innovation. Datasets published on the website have non-standardised, inconsistent contents and do not even adhere to a single format for date columns of time series data. Structured data has been published in thirteen different file formats and metadata is almost non-existent, comprising only dataset titles, descriptions, publication dates, and publisher names. Although the website is built on the CKAN architecture, the user experience regarding data discoverability and accessibility is incomparable to other, more user-centric public sector data portals across the world.

Although we attempted to mitigate this with our data processing pipeline by cleaning data, standardising file formats, and attaching metadata, the effects still propagate to the end result. If a user finds a dataset in the NDL-lite from data.gov.uk that is titled as if it may fit their needs, it is highly likely that, once they set about analysing it, they find that it is not only difficult to use but also much different to what they require. The user experience, even for AI agents, is negatively affected by issues stemming from data.gov.uk

In comparison, we note the relative ease of use of datasets from the ONS contained in the NDL-lite. Up-to-date economic statistics are easy to find, understand and use because these datasets are clearly titled, even though they lack attached metadata. The datasets are clean and standardised, making them interoperable and reusable for many kinds of research.

We consequently recommend that the NDL initiative starts by addressing data issues that will sit at its foundation. By dedicating resources to 'renovating' data.gov.uk, improving the contents of datasets and the metadata surrounding them, the government can ensure that the rest of the NDL development runs more smoothly and offers a more user-centric, AI-ready experience.

This renovation can be made easier by leveraging AI. Recent research[41] has found that modern-day LLMs can meaningfully improve titles, descriptions and other metadata to make datasets easier to find, use and re-use, and the integration of knowledge graphs to a data portal can completely alter the way datasets can be discovered.[42] As with the NDL itself, there is a large body of evidence to guide the redesign of government dataset publishing practices; the next step is turning analysis into action.

AI agents find government data hard to use, often choosing to look elsewhere

In our experimentation with AI agents during this work, we were struck by how capable they were. Agents like Gemini-CLI or Claude Code sit within a programming environment and, if given permission, can take full control of writing, editing and running code to perform what the user expects them to. With access to the internet, MCP servers, and 'plug-ins', they can do almost anything a human can, with much greater speed.

These agents employ continual trial and error to solve a problem. If an issue occurs when they attempt to carry out user instructions, whether benign or otherwise, they will seek to find a solution, trying and retrying different options until they are successful.

This behaviour has surprising implications for our research. As per the second example above, if we asked agents to use the NDL-lite to answer a research question, they would first attempt to search for datasets or unstructured data through the NDL-lite MCP server. However, when no datasets fit the bill, the agents did not give up or acquiesce that they were having difficulties. Instead, after a few attempts, they moved to search the internet for the data they needed, downloaded it from whatever sources they found – including atypical data sources like public github repositories – and answered questions that way.

This behaviour stems from the aforementioned lack of AI-readiness amongst datasets that we had within the NDL-lite. A search for crime statistics might yield, at first look, hundreds of datasets that could be useful for national-level analyses. However, on closer inspection each of

[41] Gan, L.-Y., Walker, J., and Simperl, E. (2025), 'Keywords are not always the key: A metadata field analysis for natural language search on open data portals'
[42] Gan, L.-Y. et al. (2025), 'Using Knowledge Graphs and Large Language Models in Data Discovery'

these datasets (such as the 'Crime' dataset on data.gov.uk) is titled or described misleadingly, often being statistical releases issued by local authorities that cannot be joined together easily to develop national-level analyses because of the lack of standards interlinking them.

Furthermore, the freshness of some statistics on data.gov.uk is a concern. AI agents cannot use one of the most authoritative sources of historical crime statistics in the UK we could find, the 'Crime Statistics' published by the Home Office, because the dataset was last published on the site in 2018. Notably, the more recently published version of this dataset was simply inaccessible via the ONS API, preventing its ingestion into the NDL-lite in our data curation stage.

With the utility of our prototype hampered by the quality of the data it contains, AI agents instead choose to search the internet to find the statistics they require. As a result, they rely on sources that are incomparable in authoritativeness to government data platforms, using pieces of information found online to build their responses to user prompts and never saying 'I don't know'.[43] This limits both the trustworthiness of AI agents and their overall utility, putting a downward pressure on the uptake of such innovative technology.

This paradigm can only be improved by making the NDL as robust as possible. It must be complete, it must be AI-ready, and it must be easily usable, so that researchers armed with agentic AI workflows can rely on it for datasets rather than having to search elsewhere. In doing so, the government can ensure that innovators in the UK have the ability to use public sector data to guide them and their work at the cutting edge.

[43] Majithia, N. et al. (2026), 'The CitizenQuery Benchmark: A Novel Dataset and Evaluation Pipeline for Measuring LLM Performance in Citizen Query Tasks'

Conclusions and recommendations

A successfully implemented NDL could become a permanent piece of national infrastructure, enabling the government not only to foster innovation but to capture its benefits by making public sector data easily accessible and usable for research and analysis.

The NDL-lite is a tangible prototype of what this could look like. It presents a first step towards accomplishing the aims set out in the AI Opportunities Action Plan,[44] and with our code released openly, we provide a functional foundation for the NDL initiative to build upon.

In our work, we found that making a prototype that interacts with the cutting edge of agentic AI was not as difficult as it sounds, with open-source tools and protocols being the industry standard for much of the pipeline. The difficulties encountered were largely due to the lack of AI-readiness in current public sector data repositories like data.gov.uk, with issues at the dataset, metadata and infrastructure level propagating into negative effects upon our prototype.

AI agents are extremely capable, and when confronted with poor-quality data, they tend to look elsewhere. This behaviour had important ramifications in our testing of the NDL-lite, where disjointed and fragmented datasets on public sector repositories were so difficult to use that agents chose to avoid them and instead searched the internet for statistics from other, sometimes non-governmental, sources instead.

For the NDL to actually yield the benefit it promises, the initiative must therefore contend with making its foundations – the datasets currently available on public repositories like data.gov.uk – more AI-ready. With our framework already referenced in the government's own set of principles for AI-ready data, we are confident that this process is already beginning behind the scenes at DSIT.

However, working solely behind the scenes could be an error. Without public visibility, the construction of the NDL would miss out on the direction and insights offered by key stakeholders across the technological ecosystem in the UK. The innovators, academics and researchers that use public sector data have much to say about what its future should hold.

[44] Department for Science, Innovation, and Technology (2025), 'AI Opportunities Action Plan'

By making our prototype publicly available, we ensure that there is a concrete starting point for these conversations, and in openly publishing our code, we have turned the first building blocks of the foundation of the NDL into reality. We hope that, in turn, the NDL initiative is propelled forwards, with us at the ODI supporting it along the way.

Appendix: Technical specifications

Development work established a comprehensive technical framework for the NDL-lite, designed to facilitate the aggregation, processing, and semantic retrieval of UK public sector data. The architecture is composed of four distinct layers: (1) a robust Data Ingestion and Processing Pipeline orchestrated via Dagster; (2) a centralised Data Corpus, comprising both unstructured textual data and structured tabular data; (3) a Vectorised Search Index for semantic retrieval; and (4) an Access Layer, consisting of a RESTful API, a Python client library with built-in MCP support for agentic scenarios, and a reference Retrieval-Augmented Generation (RAG) application.

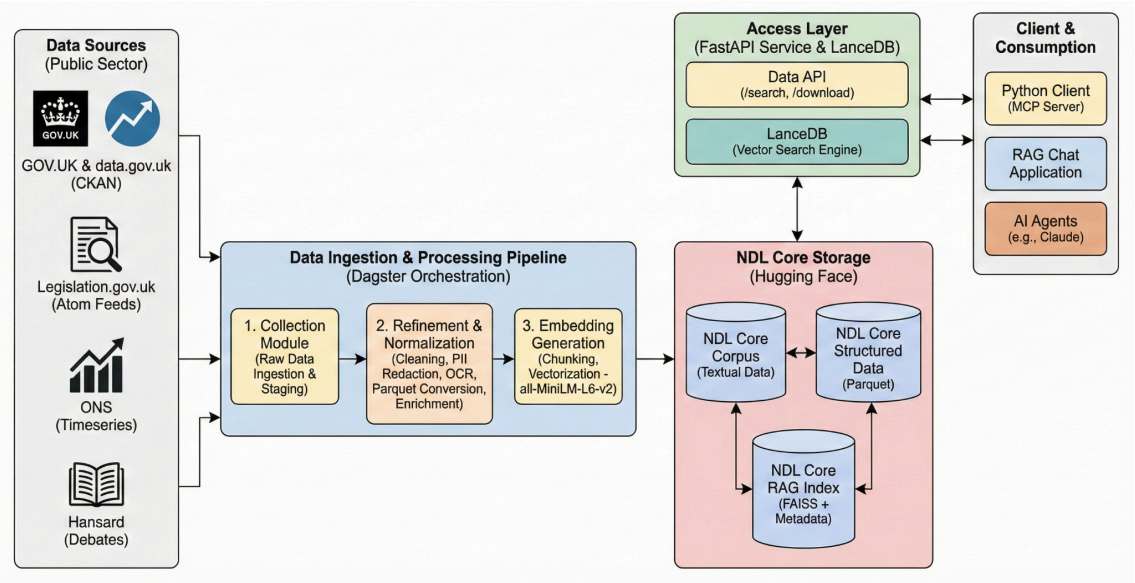

Figure 2: The NDL-lite architecture

Pipeline

(1) Data collection module

The collection module implements a series of rate-limited API clients and crawlers designed to ingest raw data from heterogeneous public sector sources. Key connectors include:

- **GOV.UK & data.gov.uk:** Harvesters for government publications and datasets, utilizing the CKAN API for dynamic category discovery and filtering by public licenses (for example, Open Government Licence).
- **legislation.gov.uk:** An ingestion engine for legal documents and statutes via Atom data feeds.
- **Office for National Statistics (ONS):** A connector fetching timeseries datasets and topic taxonomies.
- **Hansard:** A direct ingestion mechanism for parliamentary debates and Q&As.
- **environment.gov.uk:** These datasets are also published on data.gov.uk and are crawled using the same method.

Raw data is initially partitioned and stored in a staging environment to ensure isolation before processing.

(2) Data refinement and normalisation

The refinement stage transforms raw inputs into an AI-ready format through a strict processing methodology:

- **Cleaning:** The system executes automated routines for deduplication, null value handling, date times ISO 8601 standardisation, and the redaction of Personal Identifiable Information (PII) such as phone numbers and email addresses.
- **Structuring & OCR:** Unstructured formats (such as PDF) are converted to text using Optical Character Recognition (OCR) tools (Tesseract, Poppler). Text is extracted from HTML content using BeautifulSoup. Structured formats (CSV, JSON, XLSX, ODS) are standardised into Apache Parquet format to optimise I/O performance for analytical workloads.
- **Enrichment:** Records are augmented with metadata, including automatically detected language tags and EU Data Theme vocabulary tags, ensuring semantic consistency across the corpus.

(3) Embedding generation

To support semantic search, the pipeline incorporates an embedding stage using the `sentence-transformers` library. The system uses two distinct embedding databases: a chunking-based FAISS (Facebook AI Similarity Search) index to drive question answering over the NDL Core corpus, and a LanceDB to enable dataset search for a given topic.

- **Chunking Strategy:** Textual data is segmented using a `RecursiveCharacterTextSplitter` with a chunk size of 800 characters and an overlap of 100 characters to preserve context at boundaries.
- **Vectorization:** The `all-MiniLM-L6-v2` model is employed to generate 384-dimensional dense vector embeddings for each text chunk.

Data publication

The processed output is organised into three distinct datasets hosted on Hugging Face, serving as the immutable truth for the system.

- **NDL-lite Corpus:** A multi-modal dataset containing more than 142,000 records of textual data. The schema enforces strict typing, including fields for `identifier` (UUID), `title`, `description`, `source` (such as Hansard or ONS), `license`, and `token_count`.
    - https://huggingface.co/datasets/theodi/ndl-core-corpus
- **NDL-lite Structured Data:** A collection of tabular datasets (for example, from ONS and Defra) converted to Parquet. This dataset supports analytical tasks that require structured quantitative data rather than purely semantic text retrieval.
    - https://huggingface.co/datasets/theodi/ndl-core-structured-data
- **NDL-lite RAG Index**: A dedicated retrieval index containing a FAISS index, used for chat like Question Answering, and a LanceDB index to enable dataset search for a given topic. This structure guarantees deterministic mapping between the vector space and the original source text, enabling precise citation in downstream applications.
    - https://huggingface.co/datasets/theodi/ndl-core-rag-index

Access layer

To democratise access to the processed data, a FastAPI-based service layer was developed, backed by LanceDB for high-performance vector search.

Note: The API and RAG Chat application use Hugging Face free-tier spaces. Consequently, these spaces will automatically enter a sleep state after 48 hours of inactivity. They must be reactivated with a request, and this initial request may experience a longer delay due to the 'cold start' process.

Data API specifications

The API exposes endpoints for semantic search and content retrieval:

- GET `/search`: Accepts natural language queries and returns semantically relevant datasets. The back-end uses the generated embeddings to perform cosine similarity searches against the indexed corpus.
- GET `/download/text/{identifier}`: Provides streamable access to full-text records, facilitating integration with external analysis tools.

Client Library and Agentic Integration

A Python client library (ndl-core-client) was engineered to wrap the API interactions. Notably, this library implements the Model Context Protocol (MCP), allowing AI agents (like Claude and Gemini) to programmatically query the NDL Core. The MCP server exposes tools that enable agents to ‘Find UK government datasets’ directly within their native chat interfaces, abstracting the complexity of vector search from the end user.

https://github.com/theodi/ndl-core-client