LUSTRE/GLOW Workshop 6 Government Records and AI

LUSTRE
Unlocking our Digital Past with Artificial Intelligence

The sixth LUSTRE/GLOW Workshop on Government Records and AI took place at The National Archives on 20–21 April 2026. Convened by Professor Lise Jaillant (Loughborough University), it brought together researchers and practitioners from government and academia, including the UK Cabinet Office, The National Archives (TNA), University of Oxford, King's College London, and the National Archives and Records Administration (NARA) in the United States. The workshop examined how AI can improve access to government records, while making clear that existing approaches to managing digital material are under increasing strain at scale.

The scale of the challenge

Speakers emphasised that the volume of digital records has outgrown current capacity. Professor Jason R. Baron (University of Maryland) noted that fewer than 0.07% of 700 million presidential emails held by NARA are publicly accessible, while Dr James Lappin and Piers Walker (Department for Science, Innovation and Technology) highlighted the difficulty of processing email archives at scale. The gap between what is created and what can be accessed continues to widen.

Getting the basics right

Poor information management remains a central barrier. David Canning (Cabinet Office) and Dr Kelcey Swain (Cabinet Office) showed that structured, well-governed records make AI-assisted tasks such as triage and compliance possible. Where those foundations are weaker, outputs become unreliable. Robert Bath (Intelogy) illustrated this through an example from his own organisation: Microsoft Copilot told staff that his firm offered dental insurance, drawing on two documents that had nothing to do with company policy – one generated by a colleague as test data, the other from a client project. Training, governance, and information architecture were identified as essential components of readiness.

What AI can and cannot do

Applications ranged from email triage and handwritten text recognition to Freedom of Information (FOI) support and sensitivity filtering. Michael D. Thomas (NARA) described systems already operating at scale in US agencies, including one achieving high agreement with human reviewers while reducing processing time. Speakers also noted that these tools introduce new risks alongside efficiencies, particularly where outputs depend on combining material from multiple sources.

Trust, risk, and authenticity

Questions of trust ran throughout. Professor John Collomosse (University of Surrey) presented the C2PA[1] Content Credentials standard as a way of tracing provenance, while

[1] C2PA: Coalition for Content Provenance and Authenticity, an industry group that develops technical standards to record and verify the origin and history of digital content.

Dr Cassandra Kist (University of Strathclyde) emphasised that public-facing systems should communicate their limits clearly. The discussion highlighted the importance of helping users to assess reliability when AI mediates access to records.

The long view
John Sheridan (TNA) closed by drawing on Robert Solow's productivity paradox to suggest that the impact of AI will take time to materialise as workflows and organisations adapt. Across both days, participants emphasised the need for practical guidance, shared approaches, and opportunities to learn from experience. GLOW-Ready, the next phase of the LUSTRE/GLOW project will build on this by focusing on end users and exploring how AI can support access to records in ways that are workable in practice.

DAY 1

Professor Lise Jaillant (Loughborough University) opened the workshop by situating it within the LUSTRE/GLOW project's broader programme of work, noting the recent publication of the open-access scoping study, *Sifting the Digital Heap*, and a special issue of *AI & Society*. **John Sheridan (TNA)** offered a contextual welcome that framed the stakes: TNA holds a thousand years of records of public administration, and the digital heap now confronting archivists represents both the greatest challenge and the most urgent opportunity for AI application. His observation that AI offers "the best hope for managing and navigating a wildly heterogeneous and changing body of digital records" set the tone for the workshop.

KEYNOTE:
Beyond the Hype: Building Solid Data Foundations for Knowledge Transformation

David Canning (Cabinet Office) opened the keynote by grounding the AI conversation in the realities of records lifecycle management, arguing that organisations which invested in good information management before the generative AI wave are now reaping dividends through smoother adoption. The Cabinet Office's response to the digital heap has included the Lexicon[2] – an AI-assisted document review and disposal system that has enabled large volumes of digital records to be reviewed quickly – and, more recently, Project Blackbird, a proof of concept developed with the University of Salford's computer science faculty.

Project Blackbird applied BERT-based[3] natural language processing to the body text of ministerial and senior leader emails, combined with metadata analysis of correspondence networks, to rank emails by importance across 66,000 real messages from a test set of ministerial mailboxes. The findings validated several significant assumptions and

[2] Canning, D., Jaillant, L. AI to review government records: new work to unlock historically significant digital records. *AI & Soc* 40, 4433–4445 (2025). https://doi.org/10.1007/s00146-025-02221-0
[3] BERT (Bidirectional Encoder Representations from Transformers): a type of language model used for analysing text.

provided, in Canning's words, "very valuable evidence" that approximately 8% of director-level mailboxes contain material of permanent value (with 92% ephemeral); that email records exist as interconnected strings spanning multiple mailboxes; and that the meaningful unit of archival analysis is the network of correspondence rather than the individual message.

Canning used this to raise a broader question about the adequacy of the Public Records Act, arguing that the calendar year as the organising principle for record transfer is ill-suited to digital records, which cluster more meaningfully around parliamentary tenures, ministerial periods, and thematic networks.

Dr Kelcey Swain (Cabinet Office) extended Canning's argument through a demonstration of policy-as-structured-data. Compliance policies are decomposed into machine-readable formats (e.g. YAML[4] or JSON[5]) and linked to an ontology engine that builds a knowledge graph from what users tell it, grounding terminology in known ontologies such as FOAF[6] and ORG.[7] The result is a largely deterministic system – what Swain described as a 'thin AI layer' over a verifiable knowledge base – where users can describe a project in plain English and the system maps it to relevant policies, generates a structured compliance pathway, and produces traceable, auditable decisions. The AI makes no decisions itself: it functions as an interface to the deterministic system beneath it. Swain likened it to the computer in Star Trek: not intelligent, but a reliable interface for querying structured knowledge.

The Q&A covered five questions:

1. On email selection, the Cabinet Office team explained their graduated Capstone approach — adopted partly due to GDPR requirements and the risk of losing records entirely — with permanent preservation for the most senior mailboxes, AI-driven selection being developed for senior levels, and business as usual below Director level.

2. On the Lexicon, they described a platform-agnostic document review system combining semantic and format analysis to sort records for deletion or retention.

3. On the choice of Gemini, they explained that it was already within the Google estate, unlike Redbox, an earlier system never cleared for personal identifiable information because it required data to leave the estate.

[4] YAML: YAML Ain't Markup Language, a human-readable data serialisation language designed to be human-friendly.
[5] JSON: JavaScript Object Notation, a simple text format for structuring data using key–value pairs.
[6] FOAF: Friend of a Friend, a machine-readable ontology describing persons, their activities and their relations to other people and objects.
[7] ORG: a core ontology for organisational structures, aimed at supporting linked data publishing of organizational information across a number of domains.

4. An online question on the ontology engine drew out that how much human input is needed is still being worked through, and that the system records contradictory meanings rather than forcing consensus.

5. A final online question on US platforms elicited that data is hosted and processed within the EU, and that the Cabinet Office is in conversation with the team behind the UK government's recently announced sovereign AI fund.

Session 1: Rethinking Access and Interpretation

Professor Reuben Binns (University of Oxford) reframed AI as an archival threat rather than a tool, arguing that the same LLM capabilities used for redaction can also enable de-redaction – the recovery of hidden information. This risk follows directly from how LLMs are trained, masking tokens and predicting missing content. Tools combining font-aware character-width estimation with LLM candidate generation already exist, and he showed that platform guardrails can be bypassed through step-by-step conversational prompting.

For archives, the implications are significant. Technical mitigations – such as monospace fonts or full document reformatting – may reduce but not remove the risk, raising the prospect of a shift towards transferring records as 'closed' rather than redacted-and-open. Discussion also touched on whether Freedom of Information Act (FOIA) exemption categories might aid de-redaction by narrowing context, and whether sovereign AI, grounded in UK law, could offer stronger safeguards than reliance on commercial platforms.

Chris Royds (TNA) presented a proof-of-concept exploring how AI can support appraisal of large, unstructured digital collections, using the Department for Environment Food and Rural Affairs' (DEFRA) 2005 web archive (c.21,000 files). The project combined Retrieval-augmented generation (RAG)[8] for querying, LLM-based file and folder analysis, BERTopic for thematic clustering, and Apache Tika for metadata extraction, producing both a queryable system and a Power BI dashboard. This allowed archivists to identify themes, map them against folder structures, and trace decision-makers across the collection – tasks that would be prohibitively time-consuming manually. Royds noted limitations, including difficulties with UK government terminology and the relatively small scale of testing.

A proposed output, the Machine-Interpretable Selection Outline (MISO), would encode policies, legislation, and key actors in structured formats such as JSON, providing context for AI-assisted appraisal at scale. Discussion raised broader concerns about interpretation: whether summarisation risks discarding detail that different researchers value differently, and how far such systems can situate records within their wider political and cultural context.

Dr Deblina Bhattacharjee (University of Bath) presented a multimodal AI pipeline developed with the United Nations Archives and the Universal Postal Union to digitise

[8] RAG: Retrieval-augmented generation, an approach in which a model retrieves relevant documents from a specified dataset and uses them to generate responses grounded in those sources.

complex records while preserving all visual and contextual features. Combining vision models, semantic encoding, relational graph construction, and reinforcement learning from human feedback, the system captures handwriting, stamps, and structure rather than reducing documents to plain text. It delivered major efficiency gains, with UN digitisation rates increasing dramatically, and was also applied to postal records to support geospatial analysis. Bhattacharjee emphasised that the pipeline is non-generative, and highlighted the need for shared benchmarks, co-design, and international collaboration.

Discussion focused on practical constraints and scalability. Questions raised the resource demands of building and maintaining relational graphs, the limitations of AI in low-resource language contexts, and the need for wider institutional support. Participants also pointed to existing initiatives on under-resourced scripts and suggested engagement with international bodies such as the International Council on Archives as a route for extending the work.

Session 2: From Experimentation to Implementation

Dr Haibin Cai (Loughborough University) outlined key challenges in applying LLMs to government records at scale, focusing on hallucination, cost, and infrastructure. Cai presented mitigation strategies including RAG, LoRA fine-tuning,[9] multi-model voting, and mixture-of-experts architectures, and argued that distributed computing – running models across smaller machines – can reduce costs and support local data processing where sovereignty is a concern.

Discussion centred on trade-offs. Questions explored whether distributed approaches could also reduce environmental impact, and how far different architectures balance performance against resource use. Cai noted that smaller, distributed models can offer efficiency gains for simpler tasks, but that the optimal setup depends on the specific use case.

Robert Bath (Intelogy) set out what responsible Microsoft Copilot deployment requires, focusing on security, staff guidance, and records management. His central point was that AI amplifies existing data problems: poorly governed and over-shared information becomes more visible when surfaced by AI. Bath illustrated this with an example where Copilot incorrectly claimed the company offered dental insurance, drawing on unrelated test and client files. Without strong information architecture, retention policies, and content management, AI systems can generate plausible but false outputs at scale.

Discussion centred on control and oversight. Questions explored whether AI can tailor outputs based on user permissions, and how realistic human-in-the-loop models are at scale. Bath noted that while fine-grained access control is possible, it is unreliable in over-permissioned environments, and recommended simpler, site-level permissions. He also

[9] Low-Rank Adaptation (LoRA) fine-tuning is a method that updates only a small set of additional parameters in a model, allowing it to be adapted efficiently without retraining the entire system.

argued that humans cannot review everything, and must act as auditors and spot-checkers rather than decision-makers for every output.

Dr Daniel Chávez Heras (King's College London) presented the Intelligent Systems for Screen Archives (ISSA) project, a British Film Institute-funded collaboration with five regional UK moving image archives. Unlike most presentations at the workshop, ISSA works with audiovisual rather than textual records – but many of its methodological concerns directly parallel those of documentary archives. The project developed four minimum viable prototypes: semantic segmentation of long-form video using large visual-language models; place-based search and retrieval using geospatial entity linking; automated audio description for visually impaired users; and algorithmic editing interfaces for creative reuse.

A notable aspect of the project's approach was its insistence on developing tools with the archives rather than for them, using open-source models and in-house compute to ensure that the knowledge generated remains institutionally owned. Chávez Heras drew an explicit parallel with the sovereign AI debate raised earlier in the day: outsourcing AI development to external providers not only creates dependency, but means that the knowledge of how these systems work or fail is lost to the institution.

Nicholas Smith (NHS Resolution) reframed the digital records problem as one of knowledge management, arguing that records capture only a small fraction of organisational knowledge, with most remaining tacit and held by staff. To address this, NHS Resolution has developed AI-assisted knowledge capture interviews, where structured conversations with departing or specialist staff are transcribed and distilled into queryable 'knowledge agents'. These preserve expertise that would otherwise be lost, including senior leadership experience and highly specialised case knowledge.

Discussion highlighted the implications of this approach. It positions AI not just as a tool for processing records, but as a way of linking records to the knowledge behind them. This raised broader questions about how far such systems can reliably represent individual expertise, and reinforced the argument that without capturing tacit knowledge, AI-driven interpretation of records will remain incomplete.

PLENARY DISCUSSION: GLOW Scoping Study – *Sifting the Digital Heap*

The afternoon plenary brought participants together to discuss key questions from the *Sifting the Digital Heap* scoping study. There was general agreement that appraisal – deciding what to keep – should come before sensitivity review and metadata work, as reducing volume may be necessary before more resource-intensive processes begin. It was also suggested that demonstrating early reductions in backlogs could help secure organisational support and funding for longer-term AI initiatives.

On sensitivity review, discussions leaned towards a risk-based approach, with AI used to segment records into higher- and lower-risk categories, and human effort focused at the

boundary. Participants noted the distinction between AI as a knowledge layer, which can flag potential sensitivities, and AI as a decision-maker, which remains less certain. It was also observed that AI adoption may follow a trajectory of early mistakes and gradual standardisation, with a need for clearer professional guidance.

Questions of authenticity and coordination also surfaced. Participants raised concerns about AI-generated content entering the record itself, creating new provenance challenges that archives may not yet be equipped to address. Discussion indicated that a coordinated national strategy for AI in government records remains at an early stage, with a cross-sector taskforce suggested as a possible next step.

KEYNOTE:
A Use Case for Agentic AI Enhancing Access to Public Records: Challenges and Opportunities

Professor Jason Baron (University of Maryland) closed the day with a keynote that combined detailed evidence of strain within the US Freedom of Information Act (FOIA) system with a proposed role for agentic AI. He highlighted the scale of the challenge, including the volume of requests across federal agencies, long response times, and the large proportion of presidential records that remain inaccessible. These examples were used to illustrate the limits of current approaches rather than to suggest a single point of failure.

Baron outlined a multi-agent workflow in which different AI components handle request clarification, document retrieval, sensitivity review, response drafting, and audit. He emphasised that the model is not fully autonomous: human oversight remains, but at a higher level, rather than reviewing every individual decision. He also presented research on using AI to identify deliberative process privilege, suggesting that its performance compares reasonably with the variability found in human review.

Discussion raised a number of practical concerns. These included questions about system reliability if human oversight is reduced, how AI systems handle highly sensitive material, and the implications of AI-generated FOIA requests. It was noted that such requests are already emerging, with some jurisdictions considering how to respond. There was also discussion of how far the model could transfer across national contexts, with the general view that while the overall process may be adaptable, differences in legal frameworks would shape its implementation.

DAY 2

Professor Lise Jaillant (Loughborough University) again chaired proceedings, which moved from the more technical focus of Day 1 toward questions of trust, authenticity, and what it will actually take to make AI work in public archives.

KEYNOTE:
Balancing Access and Risk: Sensitivity-Aware Search in Sensitive Archives

Professor Iadh Ounis and Dr Graham McDonald (University of Glasgow) presented long-running work on sensitivity and archives, developed in collaboration with The National Archives since 2013. They traced a progression from early sensitivity classifiers to current work on sensitivity-aware search. This approach differs from traditional sensitivity review: instead of checking documents only before release, it filters results at the point of search, aiming to surface relevant material while suppressing sensitive content. This is particularly important because systems such as RAG can combine individually harmless documents to reveal sensitive information – the so-called mosaic effect.

Dr McDonald described a two-stage retrieval pipeline. An initial lightweight retriever processes the full collection, down-weighting terms associated with sensitivity so that higher-risk documents are less likely to be selected. A second-stage re-ranker then refines results, pushing any remaining sensitive material further down. A key feature of their learned sparse retrieval approach is interpretability: the model assigns visible weights to terms, allowing reviewers to understand why results are ranked as they are. Their results suggest that combining sensitivity controls at multiple stages improves both relevance and safety compared to standard search methods.

To support this work, the team developed SARA, a test collection for sensitivity-aware search based on the Enron email corpus, with added relevance judgements and sensitivity labels. While not a perfect match for government records, it provides a practical benchmark where real-world data is unavailable. They also pointed towards agentic workflows as a likely next step, with different agents handling stages of search and review. Discussion highlighted limitations of legacy datasets and the ongoing risk that sensitivity can emerge across multiple documents or queries, reinforcing the need for filtering at the retrieval stage as well as before release.

Session 3: Navigating Trust and Authenticity

Professor John Collomosse (University of Surrey) opened by noting that concerns about false information long predate AI, but that the scale and speed of generation have changed. His work centres on the C2PA (Coalition for Content Provenance and Authenticity) standard, which uses cryptographic signatures to attach a verifiable record of how digital content was created, from what sources, and by whom. Adopted by major platforms and tools, it functions as a digital equivalent of historical practices that established authenticity and ownership. He outlined how durable provenance can be maintained even when metadata is stripped, combining signed metadata, perceptual fingerprinting, and invisible watermarking to create a more resilient system.

Collomosse argued that provenance provides a foundation not only for authenticity, but also for rights management, licensing, and content discovery. He demonstrated a

prototype system that can recover provenance from images shared online and trace contributions back to original sources, enabling licensing to be distributed across multiple rights holders. He linked this to the proposed Creative Content Exchange (CCE)[10] for AI training data, suggesting that a decentralised model would better support freelancers and smaller creators as well as large institutions. The work builds on earlier archival research, showing how provenance technologies developed for collections can scale into wider digital and commercial contexts.

Dr James Lappin and Piers Walker (Department for Science, Innovation and Technology) framed the digital records challenge as one of scale without process. Lappin argued that the issue is not a blocked pipeline, but the absence of a workable one, using the example of hundreds of millions of emails with only a small fraction accessible. He outlined a capstone-style approach to reduce volume early by selecting a small proportion of accounts of potential historical interest, and then breaking these into smaller, coherent clusters to make sensitivity review manageable at a level where reviewers can be confident of coverage.

Walker described Project Kestrel, which tests this approach through three strands: generating synthetic email datasets to avoid handling real sensitive material, clustering mailboxes using unsupervised machine learning, and presenting results in a dashboard where reviewers can explore clusters and communication networks. Early work uses the Enron corpus, with plans to move to synthetic data for more controlled testing. He noted that defining what makes a useful cluster remains an open question, including how large clusters should be and how their contents can be summarised quickly enough to support review.

Professor Christopher (Cal) Lee (University of North Carolina) offered a conceptual framework for thinking about AI and digital records. He argued that documentation allows societies to revisit past states, which underpins accountability and memory, but that digital environments shift the focus from preserving objects to preserving reproducible states. This creates new challenges for curation, particularly as AI systems become part of what needs to be preserved and interpreted.

A central contribution was the idea of "dualities" – tensions where both sides are valid and must be balanced rather than resolved. Examples included just-in-case versus just-in-time processing, breadth versus depth in training data, and accuracy versus explainability. Applied to AI, this raises questions about when to build models in advance or on demand, whether to rely on general or specialised systems, and how to preserve evidence of what models contain and produce. Framing these as dualities, rather than fixed choices, provides a way to navigate complexity without forcing false distinctions.

Dr Cassandra Kist (University of Strathclyde) examined public-facing AI through the lens of museum conversational agents, arguing that archives can learn from systems

[10] On 26 January 2026, the UK government announced a pilot for establishing a Creative Content Exchange (CCE) as a trusted marketplace for selling, buying, licensing, and enabling permitted access to digitised cultural and creative assets and investment.

already deployed with real users. Drawing on case studies, practitioner interviews, and a visitor survey, she framed authenticity across three dimensions: emotional (personal connection), material (connection to objects), and factual (grounded in verifiable sources). Her focus was on collections discovery agents – tools designed to help users find material rather than create immersive experiences.

These systems tend to avoid avatars and personas, maintain strict knowledge boundaries, and use clear, transparent communication about what the agent can and cannot do. Kist described this as "transparent absence": by making limitations explicit, systems can build trust without overstating capability. Survey responses highlighted concerns about distraction from collections, loss of human interaction, hallucination, and data use. She noted that while citations can increase user trust, genuinely tracing the provenance of AI outputs remains a more complex challenge than simply linking to relevant sources.

PLENARY DISCUSSION: GLOW-Ready – Next Steps for AI in Government Records

The afternoon plenary gathered feedback for the GLOW-Ready project, which will focus on end users of archives – historians, researchers, journalists, and members of the public. Participants split into breakout groups, each selecting two questions from a prepared list covering topics from user expectations and service design to trust, limits, and what the project should prioritise. The following summarises the main points raised:

What the project should produce

- Practical examples, including honest accounts of what has not worked, which are rarely shared but often the most instructive
- Some level of standardisation: one table observed that a great deal of experimentation is happening across the sector with very little coordination
- Smaller, more regular networking groups for people working in similar areas, alongside larger events

Staff training and readiness

- Getting people confident enough to use these tools takes more than an online course
- Tools change quickly, making it hard for staff to keep up
- Organisations need people with the right skills in-house who can then help their colleagues
- Getting new software approved for use on government networks can be slow and difficult

What a helpful discovery service would look like

- Different users need different things: experienced researchers want AI to complement existing search tools; less experienced users, such as family historians, would benefit from a conversational interface that helps refine searches
- Growing public scepticism about AI is something the project will need to reckon with
- Open-source models often have built-in restrictions that may make them less effective for certain historical topics, and known biases in some models could be a problem in archival research contexts
- One table raised the question of what AI systems consistently fail to return, and whether that reflects gaps in user interest or blind spots in the models

Testing

- On how the project should test whether its ideas are useful, one table argued for depth over breadth: fewer, more thoroughly evaluated pilots
- Success criteria defined in advance
- Testing conducted with realistic data in real working environments

User expectations and project outputs

Dr Kelcey Swain's table addressed questions on user expectations and project outputs. On users:

- External users want accountability, but fewer than one in a thousand records are currently accessible on request
- More could be done at the point of document creation to embed metadata automatically, reducing the retrospective burden
- Tiered access proposed for records held before they reach The National Archives: curated access on request, free access for certain parts, restricted access elsewhere
- For people working inside government, the goal would be finding archived material as easily as running a search

On project outputs, the table identified **three priorities**:

- A secure environment for cleared researchers to work with government material
- A map of relevant expertise across government and academia, and what questions remain unanswered
- A way of consulting archive users about what they want to see, noting that the current approach captures only a thin slice of senior decision-making with almost nothing about how the civil service functions day to day

KEYNOTE:
To Publish and Declare – Fulfilling our Foundational Promise of Transparency in Government

Michael D. Thomas (NARA) joined the workshop online to discuss the management of classified government records, describing it as a "parallel universe" to public archives with similar challenges of scale and access, but shaped by national security constraints. He opened with the Dunlap broadsides of the Declaration of Independence, arguing that the act of publication – rather than signing – marked the first declassification, establishing a long-standing tension between secrecy and the obligation to make records public.

He outlined two current applications of AI in US government practice. At the State Department, machine learning systems trained on past classification decisions are used to sort diplomatic records into those suitable for release, those requiring retention, and those needing further review, achieving high agreement with human reviewers while reducing processing time. At the Department of the Air Force, the Battery RAM project focuses on classification at the point of creation, using security classification guides to build a structured knowledge base to support or automate decisions earlier in the lifecycle. Both examples were used to argue that consistent, structured decisions upstream reduce the burden of review downstream.

Discussion focused on implementation and transferability. Questions addressed how far these approaches depend on the availability and quality of prior decision data, whether similar systems could be developed in contexts with less consistent classification practices, and how far external pressure – rather than technical readiness – drives investment in such systems. A related question raised the implications of recent legal interpretations around presidential records, pointing to some uncertainty about how record ownership and classification responsibilities intersect in practice. There was also reflection on the relationship between classification and declassification, and whether improving one side of the process meaningfully reduces pressure on the other.

KEYNOTE:
Productive Digital Archives

John Sheridan (TNA) closed the workshop with an assessment of where archives stand with AI. He framed the current moment through an AI-generated 'toolbox' image: some tools are clearly powerful, others obviously useless, and the challenge is working out which is which through use rather than assumption. Drawing on Shannon Vallor's *The AI Mirror*, he argued that large language models represent a fundamentally different kind of intelligence to humans – built from arithmetic upward rather than evolved cognition – and that the tendency to anthropomorphise them leads to misplaced expectations about both their capabilities and their limits.

He placed this within a longer history of technological change, invoking Robert Solow's productivity paradox to argue that visible technological advances do not translate

immediately into measurable gains. AI's impact, he suggested, will depend less on the tools themselves than on the slow redesign of workflows, evaluation methods, and organisational structures. His comparison between Microsoft Copilot and Claude Code illustrated this: the latter appears transformative partly because software engineering provides built-in mechanisms for testing and verification, whereas information management lacks comparable feedback loops.

Sheridan identified the Model Context Protocol (MCP),[11] developed by Anthropic, as a promising way forward. By connecting chatbots to structured, domain-specific data sources via APIs, MCP allows systems to perform far better within bounded contexts than as general-purpose tools. Experiments at The National Archives – including linking models to legislation data – showed clear improvements, and similar approaches are emerging elsewhere. He also pointed to successful internal evaluation work, such as using LLMs to extract legislative powers and duties at scale with high accuracy, while stressing that such results depend on tightly defined, testable tasks.

He ended by emphasising that progress will be incremental and collaborative. Many organisations are working on similar problems in isolation, often unaware of parallel efforts, and the biggest gains may come from sharing approaches, including failures. Workshops of this kind, and networks like GLOW, were presented as essential spaces for building that shared understanding and avoiding duplicated effort.

THEMES AND CONCLUSIONS

Several themes recurred across the workshop:

- **The manual model is under increasing strain.** Speakers and participants pointed to the scale of born-digital records, email archives, FOI backlogs, and sensitivity review as evidence that existing processes cannot simply be extended by adding more human labour.

- **AI readiness depends on records and information management readiness.** Poorly structured, poorly governed, or poorly understood records can produce unreliable outputs, while better metadata, retention practice, permissions, and information architecture make AI-assisted triage, appraisal, compliance, and discovery more viable. Enrichment at the point of creation was also highlighted as a way to reduce downstream burden.

- **AI introduces risks alongside efficiencies.** These include hallucination, de-redaction, over-redaction, inappropriate disclosure, and the mosaic effect, where individually non-sensitive records can become sensitive when combined. Several contributions also showed that AI systems may surface irrelevant, outdated, over-shared, or poorly governed material as though it were authoritative.

[11] Model Context Protocol (MCP) is a standard that enables AI models to connect to external data sources and tools, providing structured context that they can use to generate more accurate responses.

- **Human oversight must be structural.** A recurring tension across the workshop was between the principle of human-in-the-loop review and the practical difficulty of applying it at scale. Risk-based models therefore place human attention on boundary cases, auditing, quality sampling, and escalation rather than every AI output.

- **Knowledge management is infrastructure.** NHS Resolution's knowledge capture interview programme pointed to a dimension of records AI that is easy to overlook: the tacit knowledge needed to interpret records is not in the records themselves. Capturing that knowledge before it is lost is as important as digitising the documents.

- **Institutional ownership of AI knowledge matters.** The ISSA project's emphasis on open-source, in-house development, and the recurring theme of sovereign AI reflected concern that reliance on commercial providers may externalise learning about how these systems succeed and fail. That learning is itself a form of institutional knowledge that archives need to retain.

- **Trust depends on provenance and transparency.** AI outputs are more reliable when users can trace the sources that informed them, understand the limits of the system, and inspect the evidence behind a response.

- **Organisational readiness remains a constraint.** AI adoption depends on staff training, governance, realistic pilots, technical confidence, and the redesign of workflows, not simply access to new tools.

- **Evaluation remains unresolved.** Assessing AI outputs requires expert input, which becomes impractical at scale. This strengthens the case for shared evaluation methods, open datasets, and collaborative testing.

The two-day workshop concluded with an invitation to join the **GLOW mailing list** and a commitment to share a full written report drawing on both days' discussions and the feedback gathered in the table exercises. GLOW-Ready was also positioned as the next step, with a focus on practical guidance, user-centred testing, and shared learning across government, archival, and research communities.

GLOSSARY

BERT (Bidirectional Encoder Representations from Transformers**)**, a type of language model used for analysing text.

C2PA (Coalition for Content Provenance and Authenticity), an industry group that develops technical standards to record and verify the origin and history of digital content.

FOAF (Friend of a Friend), a machine-readable ontology describing persons, their activities and their relations to other people and objects.

FOIA (Freedom of Information Act), the US law that gives the public the right to access federal government records.

JSON (JavaScript Object Notation), a simple way of organising data as labelled pieces of information (key–value pairs) in plain text.

LLM (large language model), an AI system trained on vast amounts of text that can understand and generate human-like language.

LoRA (Low-Rank Adaptation) fine-tuning is a method that updates only a small set of additional parameters in a model, allowing it to be adapted efficiently without retraining the entire system.

MCP (Model Context Protocol), a standard that enables AI models to connect to external data sources and tools, providing structured context that they can use to generate more accurate responses.

ORG, a core ontology for organisational structures, aimed at supporting linked data publishing of organizational information across a number of domains. https://www.w3.org/TR/vocab-org/

RAG (Retrieval-Augmented Generation), an approach in which a AI model retrieves relevant documents from a specified dataset and uses them to generate responses grounded in those sources.

YAML (YAML Ain't Markup Language), a human-readable data serialisation language designed to be human-friendly.