John Sheridan is the Digital Director at The (UK) National Archives. He provides strategic direction to the archive's work as a digital archive as well as overseeing its digital services. John's academic background is in mathematics and information technology, with a degree in Mathematics and Computer Science from the University of Southampton and a Master's Degree in Information Technology from the University of Liverpool.

John is passionate about computers, archives, digital preservation and open standards. He is a former co-chair of the W3C's e-Government Interest Group and serves on the UK Government's Open Standards Board, which sets data standards for use across government. John was an early pioneer of open data and remains active in that community.

[Neide Alves Dias De Sordi] *Could you comment on your professional background and the work you did it in the digital field at The National Archives (TNA) of the United Kingdom?*

[John Sheridan] Like many people working in digital archiving or digital preservation, I am an accidental archivist. I have always been fascinated by computers, since I was quite young. At University, I studied mathematics and computing and I went on to gain a Master's degree in Information Technology. I was very excited about the internet, and worked for two different start-up companies through the Dot-Com Boom in the late 1990's and early 2000's, developing software. I joined the Civil Service in 2004, and my department was merged with The National Archives in 2006. I never knew I wanted to work in an archive, but it has been my working life for fifteen years. I've really enjoyed my time here, so far.

[Neide Alves Dias De Sordi] *What were the biggest challenges and opportunities you faced in relation to the theme of data and records? What are the most significant actions you implemented in the digital field of the TNA?*

[John Sheridan] Archivists have known for a few decades that the shift from tangible (parchment and paper) records to intangible (digital) records is a big deal. In some ways there is less risk to digital records – we can make multiple copies, and store copies of the records in different places, quite easily. In some ways there is more risk.

Stepping back, every part of the archive's work – from appraisal, selection, sensitivity review, transfer, preservation and access – is more complicated for digital records. Whilst technology is involved, the big issues in digital archiving are not fundamentally technology issues. For example, how the archive makes decisions about what type of access it provides. One option is open publication on the web, but that's a bold step – and hard to reverse in the case the record contains information that should never have been published. Perhaps the archive needs to "gradate" access in various ways, to manage that kind of risk.

We see the challenge of archiving digital records as primarily being a challenge to archival practice. Together, we need to develop new archival practices for the digital age. At The National Archives in the UK, we've talked about being the disruptive digital archive. We have set out to actively develop new archival approaches.

In terms of significance, I hope we've achieved several things over the last few years. The framing of digital preservation as risk management over time

– not a new idea in the field, but a sharpening of focus for us at TNA, and the development of approaches that seek to empirically measure the difference the archive makes in terms of lower risk. I have in mind here the work we have done around digital preservation risk modelling, in particular the DiAGRAM (Digital Archives Graphical Risk Assessment Model) model and tool. Another area is through having a focus on delivering value to users of the archive and also back to the record creators. We have taken some big strides in terms of making the transferring process for digital records easier.

There is always more to do though, especially around access. We are also re-developing our cataloguing, with a new data model for our whole collection – physical and digital. These projects are difficult and require commitment, persistence and patience. By the time I finally leave The National Archives, whenever that might be, I hope to have made a lasting difference.

[Neide Alves Dias De Sordi] *TNA was one of the pioneer national archives in the management and preservation of web content of government agencies. Could you comment on the design and implementation of this project? Does the TNA use private cloud resource to store government web pages? What solution would you recommend to the Arquivo Nacional [National Archives of Brazil] to develop a similar project for archiving web pages?*

[John Sheridan] I think it is important that archives capture records from the web. TNA was right to select UK government websites as being worthy of permanent preservation. Publication has a really significant impact on the likely historical value of the record.

Web archiving is a different kind of digital archiving, in the sense that the web archive captures and compiles the record. From a technical standpoint, a web archive is not much more than a series of HTTP requests issued by the archive, and the responses that were received at the time. It's the replay software that brings the web archive to life – giving the user a sense of moving around an archived website.

As a web archive, we only archive government information. Whilst our scope is much smaller than the Internet Archive or national libraries' web archives we add enormous value by capturing swathes of content that other web archives are likely to miss. That's mainly because of the effort involved. We do our best work in capturing hard to archive content. We also try and capture records across channels, from a variety of social media services.

To a national archive looking to move into web archiving, I would say "go for it"! There is a great set of tooling available, cloud technologies make it easier

to operate at scale, and the community is very supportive. In terms of tools, then I would recommend a mixed approach – combining traditional crawling techniques using Heretrix, with other methods using tools like Browsertrix. The whole Webrecorder family of tools is amazing, and well worth investigating, especially by archives starting to archive web content for the first time.

[Neide Alves Dias De Sordi] *Could you comment on the guidance given by the* TNA *to government agencies related to the management and transfer of data and open data (open government)? What would be the methodological approach of records management to data and open data? How is data preservation and access management carried out in* TNA?

[John Sheridan] This is a big topic. Digital records can be of a wide variety of things, including structured and semi-structured data. We have been very clear to transferring departments that datasets can be public records worthy of permanent preservation, and we have made sure we have the capacity to receive, preserve and provide access to those records. The UK government has been committed to open data for over ten years now, and we have been able to capture many published datasets into our web archive. Indeed, data archiving is an important part of our web archiving activities. We have an incredibly comprehensive and deep record of *https://data.gov.uk* over the last ten years for example, including tens of thousands of the open datasets that have been published.

In terms of techniques for preserving data, like many other archives we use the Software Independent Archiving of Relational Databases (SIARD) standard. This is an open format developed by the Swiss Federal Archives, for archiving relational databases. We like that it is vendor-neutral, taking advantage of other widely used standards like XML and SQL:1999. One of the big issues with archiving data is what experience the user should have in terms of access. As end users, most of us don't work with SQL directly, we work with computer systems that are higher up the technology stack. There are some difficult questions for the digital archive to navigate, in terms of preserving data and queries at a database level, through SIARD, or trying to capture and preserve something more of the user's experience. This is now easier to do, with systems that run in a web browser – because we can deploy our various web archiving techniques. Packaging and emulating whole software systems is incredibly difficult – and perhaps creates as much long term preservation risk as it solves, in terms of providing an authentic user experience.

[Neide Alves Dias De Sordi] *Blockchain technology has been used to store and preserve metadata related to digital records and data. Please comment on the Archangel (Trusted Archives of Digital Public Records) Project developed by TNA, Open Data Institute (ODI) and the University of Surrey, and the applicability of blockchain and artificial intelligence (AI) technology to digital records and information security in general and to government agencies and public archives in particular.*

[John Sheridan] One of the joys of working at The National Archives is that we are an independent research organisation. We were naturally interested in blockchain as a record keeping technology. It was an obvious step to want to research how archives might use blockchain. That is what we did in the Archangel project with the University of Surrey and the Open Data Institute.

Our application involves using a blockchain to demonstrate that digital records have not changed (immutability). We developed a viable model for doing this in the context of a digital archive. Without going into too much technical detail, we found that a "proof of authority" model was a better approach to running an archival blockchain than "proof of work". We also experimented with developing content hashes – signatures for video content that can be used across different encodings of the content.

It is going to be increasingly important for archives to demonstrate the authenticity of the digital records we hold – particularly as techniques to synthesise digital content become more common place. In Archangel we showed how a blockchain solution might help with that.

[Neide Alves Dias De Sordi] *TNA has developed a collaborative digital preservation program. How can TNA's collaborative technology options support digital data and documentary heritage preservation risk management? How the community has been involved in the Program?*

[John Sheridan] Digital preservation is an international team effort. Memory institutions around the world rely on each other, to develop and sustain the capabilities needed to preserve and provide access to digital records. A good example of this collaboration is the registry of file format signatures we help to maintain at The National Archives, called PRONOM. There's been over a hundred different organisations around the world who have contributed to that vital resource.

Collaboration is not just about tools though. We learn from each other, through the community. I'm very honoured to be a member of the executive committees of the Digital Preservation Coalition (DPC) – which started

life in the UK, but now has a flourishing international membership, and also the Document Lifecycle Management (DLM) Forum, which brings together European national archives. Both organisations strive to bring people to together, to listen and learn from one another.

Similarly our work to develop a model of digital preservation risk, DiAGRAM. We were careful to incorporate established standards and approaches into our model – things like the National Digital Stewardship Alliance (NDSA) levels of digital preservation. We also found ways of eliciting the judgements of a range of experts, bringing those into our model.

From standards, to tools, to data, to models, to know-how, we are part of, and rely on an international community of archivists.

[Neide Alves Dias De Sordi] *Does TNA already use AI regularly to assist the English government agencies in the classification, evaluation and disposal of data and records? Could you comment about this project?*

[John Sheridan] Firstly, it is important to know that we are not responsible for deciding what records should be selected for permanent preservation. That is the responsibility of the record creators, in government departments, who make those decisions under our guidance and supervision.

That said, it is quite widely accepted that appraisal and selection decisions are going to be supported by AI tools. We think AI has a lot to offer and we are keen to support adoption. To help move that along, we have undertaken an extensive survey of the capabilities of current generation of AI tools for selection and sensitivity review, and have published the results, in a report "Using AI for Digital Records Selection in Government". This gives guidance for records managers on using AI, based on an evaluation of the current solutions on the market.

It is clear that AI cannot replace the expertise of Records Managers but can be a useful tool to help deal with the scale of digital records collections in government departments. Records Managers' knowledge of the records in their custody is going to be essential for any of the current AI approaches to be used successfully.

You can read more at *https://cdn.nationalarchives.gov.uk/documents/using-ai-digital-selection-in-government.pdf*.

[Neide Alves Dias De Sordi] *What are the challenges to archival professionals today and in the future in relation to the management, preservation and access to digital records, especially government data? What topics archivists still need to study to better understand the digital universe?*

[John Sheridan] I'm not a trained archivist, but rather a mathematician and computer scientist working in an archive – so in some ways I'm poorly qualified to speak to this issue. That said, of course I'd like to see archival education shift firmly from physical to digital records.

In a national archive setting, then archivists work as part of multi-disciplinary teams. In my view, this is the best approach – where archivists work hand in glove with user researchers, designers, developers, data scientists and so on. It means the archivist can concentrate their contribution on the core business of archiving – preserving digital records, retaining intellectual control – with a firm grasp on user's needs and expectations. There is a great importance in putting user needs first.

That said, on a technical side, it is becoming increasingly clear that archivists and records managers will need to strengthen our knowledge and skills around data analysis and machine learning, if we are to use these tools successfully.

The best way into this, I'd suggest, would be to learn some maths – particular linear algebra (vectors and matrices), and statistics (particularly Bayesian statistics and probability). And also, some computer science (Turing's ideas of computability, Shannon's information theory) and perhaps some computing history. The stories around how the different programming languages were developed or around the history of operating systems like Unix are fabulous. Learning some computing history can give a real sense for how the layers and layers of technology we see in a modern digital system came about.

[Neide Alves Dias De Sordi] *Do you believe the cooperation between the TNA and the Arquivo Nacional [National Archives of Brazil] can help the later in the developing of its digital projects?*

[John Sheridan] It has been an honour and a privilege to be involved in developing the collaboration between the National Archives, of Brazil and The (UK) National Archives.

There is so much still to do – around data and AI, selection, transfer, access, preservation and risk modelling, as well as the archivist's historic mission of retaining intellectual control. We learn from each other. I am looking forward to building our partnership and to working together more closely in future.

Entrevista realizada por Neide Alves Dias De Sordi, ex-diretora geral do Arquivo Nacional (2019-2021), em novembro de 2021.