ARCHIVES AND RECORDS
2021, VOL. 42, NO. 1, 58–78
https://doi.org/10.1080/23257962.2021.1873121

Routledge
Taylor & Francis Group

Safeguarding the nation's digital memory: towards a Bayesian model of digital preservation risk

Martine Barons [a], Sidhant Bhatia [b], Jodie Double [c], Thais Fonseca [a], Alex Green [d], Stephen Krol [e], Hannah Merwood [d], Alec Mulinder [d], Sonia Ranade [d], Jim Q Smith [a], Tamara Thornhill [f] and David H Underdown [d]

[a]Applied Statistics and Risk Unit, University of Warwick, Coventry, UK; [b]Monash University, Subang Jaya, Malaysia; [c]Special Collections, University of Leeds, Leeds, UK; [d]The National Archives, London, UK; [e]Monash University, Melbourne, Australia; [f]Corporate Archives, Transport for London, London, UK

ABSTRACT

Preservation of digital material is a challenge for which many archives feel underprepared and ill equipped. The National Archives (UK) has been working in collaboration with statisticians from the University of Warwick and partners from across the UK archives sector to develop a decision-support system which quantifies the risks involved in digital preservation. Through interdisciplinary collaboration, this partnership has developed an interactive tool for managing risks to digital material, based on a Bayesian statistical network. The tool provides archivists with a different way of thinking about digital preservation, supported by an evidence base they can use to advocate for action. The project illustrates the potential benefit of a collaborative approach, combining insight from different disciplines.

ARTICLE HISTORY
Received 22 July 2020
Accepted 13 November 2020

KEYWORDS
Digital preservation; risk; collaboration; Bayesian network; decision-support system

Introduction

Project background

Digital archives, and the materials held in them, are rich, complex and fragile. They are under threat from rapidly evolving technology, outdated policies and a skills gap across the archives sector. To preserve this heritage for future generations, archivists must understand and navigate a wide and ever-shifting landscape of risk. This is a challenge which no single archive is currently equipped to address, and which can be met only through sharing our knowledge, embracing new methods and learning from other disciplines.

The National Archives believes we are moving from an era of relative stability in archival practice focused on predominantly analogue collections, to one of continual change.[1] Each new generation of digital technology gives rise to a new set of risks with

CONTACT Alex Green alex.green@nationalarchives.gov.uk

Throughout this paper, digital material refers to anything stored within a digital archive. This includes born-digital records, digital surrogates created from analogue records, and digitized records where the digitized version of analogue material becomes the record held by the archive. Digital record is used where the nature of the material as a record is particularly important.

which digital archives (i.e. those archives holding only digital material or both analogue and digital material) must keep pace. The Archives Sector Workforce Development Strategy published in October 2018 identified the need for an increase in digital skills, and a subsequent survey of UK archives in 2019 provided further evidence of the current skills shortage.[2] There is also a widening gap between the resources we have and the resources we need. The techniques that could help close this gap — structured risk management, supported by sound evidence and robust statistics — are currently beyond the reach of the archives sector.

At the same time, the sector is facing a tipping point in the volume of digital material to be preserved. There is an urgent need for tools that will enable archivists to understand and explain risks and act to manage them.

This lack of skills and tools threatens the digital archive in several ways:

Risks to survival (Preservation risk)

For most of our history, archivists have worked with tangible records. With good care, most physical records are highly resilient. Their static nature and slow rate of change gives rise to predictable failures. Digital records are different. They are made of fluid and fragile data. They are not simply documents, but richer and more diverse types of content: threaded discussions using web-based tools, video, websites, structured datasets, and computer code. Digital records are often composite objects, potentially with multiple creators and owners.

Digital records rely upon short-lived software and hardware for their survival and will rarely last even a decade without intervention[3]. In this volatile and vulnerable environment, change is continual and the interdependence between software, hardware, data and code is increasingly complex. Threats are varied and evolve rapidly, outpacing our ability to understand and manage them. This leaves the survival of our digital heritage at great risk.

Risks to context and provenance

Digital records are intangible and invisible. Creators easily amass chaotic 'digital heaps' of unsorted information that would have been impossible to ignore in physical form. This disarray removes records from their original contexts and can reduce a coherent, usable record to isolated units of data that lose much of their meaning. This is a threat to future access, trust and value for digital collections.

Risks to transparency, trust and inclusion

The digital archivist is part of the creation story of the record. Despite a professional emphasis on neutrality, whenever archivists act to capture, contextualize, preserve and present the record, archival processes and biases inevitably creep into the story these records tell. This is exacerbated for digital records, which require deeper and more frequent intervention than analogue formats.

A failure to recognize and articulate this risk arises from a narrow focus on risks inherent in technology, to the exclusion of the impact of our own actions. Unless we can include these risks alongside other, more easily quantifiable threats, digital archives risk losing touch with the reality of our society's changing values and diverse culture.

Policy risk

Preservation policy across the UK digital archives sector is heavily informed by a set of models, standards, and processes which tell digital archives how to operate and define criteria for assessment or accreditation.

Even the best-resourced archive cannot implement every conceivable measure to reduce risk. We must all prioritize and make pragmatic decisions within the resources available to us. However, our standards are rigid and prescriptive by definition. They cannot measure risk or aid prioritization but represent an idealistic benchmark. To be compliant, an archive must achieve all that the standard specifies. This often leads to wasteful investment in actions that are not relevant in a particular environment. At the other extreme, we see that small local or community-based archives, operating within very limited resources, are at risk of not being perceived as trustworthy custodians of their own heritage because they cannot achieve full standards compliance.

Current digital preservation standards focus on the technical challenge of building and sustaining a digital repository. This requires a significant investment in infrastructure which, in turn, drives archives to prioritize technological solutions over work to address other, perhaps more pressing, threats. Our over-reliance on a standards-based approach in institutional digital preservation policies is not only limiting but potentially harmful — in effect, our policy environment becomes a source of risk to the archive. As our resources become increasingly constrained, we must equip archives with tools that help target our investment to drive better, evidence-based outcomes for the material with which we are entrusted.

Digital preservation as risk management

The idea of treating digital preservation as a risk management process is not new. In his 1996 report 'Preservation in the Digital World', Paul Conway observed:

> Organizing for preservation in the digital world is not, first and foremost, a search for process efficiency, as has been the case with traditional preservation, but rather an ongoing process of risk management, where the cost of digital file migration is judged against the cost of failure to preserve the files in terms of the patrons who need the information. The stakeholders in this organization extend well beyond the bounds of a preservation department or the administration of a library or archives to encompass technology specialists, marketing experts, and commercial vendors.[4]

There are different approaches to managing risk. As a UK government department, The National Archives follows the principles of the Treasury Orange Book.[5] More generally, ISO 31,000 establishes risk management procedures for use in business, and papers by Barateiro et al. examine this in the context of digital preservation.[6]

Vermaaten, Lavoie and Caplan provide a different digital preservation risk assessment model.[7] The 'SPOT model' defines six digital desirable preservation outcomes (*availability, identity, persistence, renderability, understandability*, and *authenticity*). The paper also usefully reviews other risk-based approaches to digital preservation. In particular, the authors note that several previous models are limited to particular subsets of threats, and that some threat categories tend to be described at a more granular and detailed level than others.[8]

Of the prior work reviewed by Vermaaten, Lavoie and Caplan, a key step is a paper by Rosenthal et al. which distinguished the bottom-up, risk-led approach, from the top-down requirements of the OAIS Reference Model and related standards.[9] Linking the two approaches are audit and certification schemes such as DRAMBORA, CoreTrustSeal (formerly Data Seal of Approval), NESTOR Seal and ISO 16363 (previously TRAC). In effect, these provide a set of mitigations against digital preservation threats though, as noted by Vermaaten, Lavoie and Caplan, they do not always make clear which threats a mitigating action is intended to address. DRAMBORA is the most structured, and its authors also refer to its bottom-up nature, suggesting that it should be used as a means of driving continuous improvement for the archive.

Of later work, perhaps the most widely adopted are the NDSA Levels of Preservation, first launched in 2013 and recently revised.[10] Since 2018, the UK Archive Service Accreditation standard has also required applicants to include an assessment against the levels.[11]

As a result, there is well-established and widely used guidance available to help identify the risks to digital archives, and mitigating practices are commonly in place. However, there is no guidance on how to quantitatively assess the risks and tailor action to the actual level of threat. This is vital if archivists are to recommend proportionate action and be seen as credible advocates for digital preservation.

The lack of a quantitative aspect to the existing general models makes it difficult for archives to determine which potential mitigation will deliver the greatest impact in reducing risk to their digital material, or conversely, to determine which risk (across all possible threat areas) is actually the most severe of those currently facing the archive. Meanwhile, the specific models may lead us to believe that these areas are the most pressing risks simply because these detailed models exist.

Any mitigating actions archivists take will incur cost, so decision makers must understand the expected benefits of these actions in order to assess them objectively. Without quantitative evidence, it is also harder to build an effective business case or advocate to funders and stakeholders who often have competing priorities. Securing resources for archives may be more challenging than for other sectors as the impact of inadequate preservation practices may not be understood until far into the future, so there is often no immediate or short-term evidence from which the decision maker can either draw reassurance or learn.

Aims and objectives

The *Safeguarding the Nation's Memory* project aims to help archivists manage digital preservation risk through the creation of a new quantitative risk management framework. The framework is expressed in the form of an interactive web-based decision support tool known as DiAGRAM, the Digital Archiving Graphical Risk Assessment Model. The project brings established statistical methods into the digital heritage sphere for the first time through close collaboration with specialists working in this field. In addition, a partnership with archives from across the UK has allowed the creation of a structured evidence base through pooling our collective evidence and experience. This highly collaborative cross-disciplinary approach with the vision of sector-wide benefit is key to this work and will be discussed below alongside the project's aims, methods and outcomes.

Diagram will:

- Improve users' understanding of the complex digital archiving risk landscape and the interplay between digital archiving risk factors

By design, the statistical techniques used to build the model reflect the conditional dependencies between digital preservation threats. As users spend time exploring scenarios and policy options within the tool, they will begin to explore this complex interplay themselves and be able to take a more holistic view of risk events.

- Enable archivists to compare and prioritize very different types of threats to the digital archive

Unlike models that have come before, Diagram will allow users to make a direct comparison of the impact of risk events of very different types, such as storage failure, insufficient metadata, loss of integrity and physical disasters. Through this, we hope to enable archivists to make better informed decisions, achieve better value for money and support them in advocating for targeted action to manage specific digital preservation risks in their own operational contexts.

- Operate even where there is limited data or imperfect evidence

There has been very limited quantitative research to date on the likelihood and impact of digital preservation risks and the long timeframes concerned add to the difficulty of conducting such research. Where evidence does exist, it often relates to measurable technological events such as storage media failures and can be very specific and detailed, making it difficult to generalize for a more holistic approach. The statistical methodology used to develop DiAGRAM allows unknown probabilities to be elicited from subject matter experts in a structured and robust way, filling the gaps where there is little or no data currently available. This overcomes a significant barrier to applying more rigorous approaches to quantifying and managing digital preservation risk and opens the way to routine application of statistical methods to support decision making in digital archives.

- Enable a wider range of people to be involved in heritage

As a broader outcome, this project aims to build awareness of digital preservation and widen access to tools for assessing threats and planning action. In addition to working directly with project partners from UK archives spanning local government, corporate and academic organizations, the project is delivering a wide programme of engagement with national and international archives.

Uniquely, the project will also build a collaboration between the heritage sector and statisticians. This will raise awareness of our digital heritage with a group that does not usually engage with archives, increase awareness of the archives sector as an employer of statistical experts and create the potential for further successful collaborations to improve the heritage sector's access to new techniques.

Materials and methods

Initial exploration of Bayesian modelling at the national archives

The concept of a Bayesian model for Digital Preservation Risk was first expressed in The National Archives' 2017 Digital Strategy.[12] Initial exploration of this idea followed the principles outlined by Fenton and Neil.[13] These principles, alongside worked examples and experimentation with a trial version of their AgenaRisk software, enabled the creation of a small proof-of-concept model which demonstrated the feasibility of this statistical approach for exploring digital preservation risk.

Wider reading suggested potential parallels with more established applications of Bayesian risk modelling techniques. For example, previous research into modelling the condition of railway bridges and determining an appropriate frequency of inspection[14] could be applied in the digital preservation sphere to model the deterioration of a digital collection over time and determine an appropriate frequency for fixity checking.

It rapidly became clear that specialist statistical knowledge would be required to make this vision a reality. Existing connections with the Alan Turing Institute brought The National Archives into contact with Turing Fellow Professor Jim Q Smith and, through him, the University of Warwick's Applied Statistics & Risk Unit.[15]

Integrated decision support systems and Bayesian networks

Digital archivists often rely on experts with disparate fields of expertise when making policy choices in their complex, multi-faceted, dynamic environment. For those wishing to make evidence-based decisions which will best support risk reduction, one of the problems faced is how to access the information and evidence they need and how to combine it to design and evaluate alternative risk management policies and actions.

An Integrated Decision Support System (IDSS) takes inputs from panels of domain experts and links them to allow a decision-maker to score each of the candidate policies she is considering, based on her measure of success — called 'utility'. This utility can have many attributes, for example, it could combine the risks of losing different formats of record with the reliability of the preserved record and the cost of undertaking the policy intervention. The IDSS aids decision makers in understanding a problem by providing a clear evaluation and comparison of the possible options available. It can be built up piece-by-piece to combine expert judgement with data for individual subsystems, and then combine these to create a full inferential procedure able to represent even very complex systems. Each subsystem in turn may be underpinned by complex models and large data streams. Relevant information from each subsystem may be provided either to the decision maker or to another subsystem which relies on it.

The IDSS paradigm can be used with a range of different overarching statistical models. The model which best captures the complex interaction of factors in the digital archive landscape is the Bayesian Network.

A Bayesian network represents the subsystems as nodes and their dependencies as arcs (arrows, Figure 1). Formally, a Bayesian network consists of a directed acyclic graph and a set of independence statements, which tell us where the arc should be present and absent.

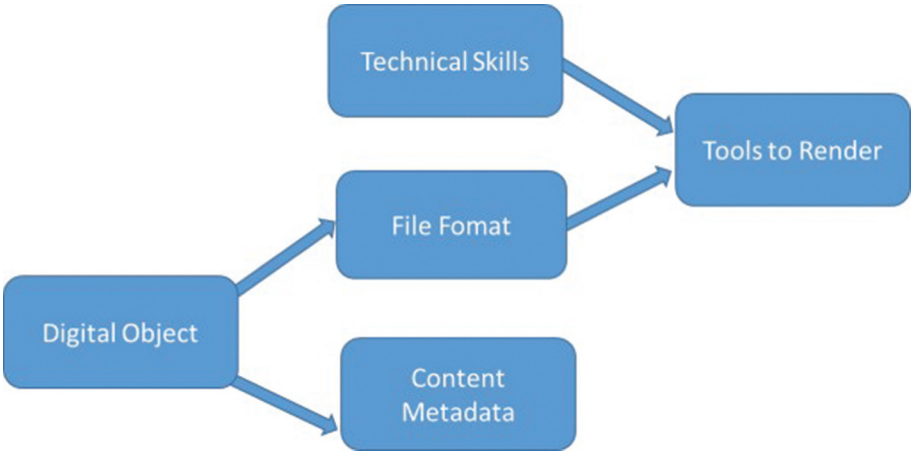

Figure 1. Bayesian Networks are flexible statistical models which accommodate the complex relationships between variables of a system (nodes). The arrows (directed edges) show direction of influence. In this example, the representation asserts that 'Tools to Render' depends on 'Technical Skills' and 'File Format', and the 'File Format' depends on the type of 'Digital Object' which also influences the 'Content Metadata'.

Expert elicitation and the IDEA protocol

Where good quality data is not available, a 'structured expert judgement elicitation' technique can be applied.[16] This well-established technique derives the required data by aggregating estimates from a panel of experts.

The IDEA protocol is one of several useful methods for structured expert judgement elicitation. The acronym IDEA arises from the combination of the key features of the protocol: it encourages experts to *Investigate* and estimate individual first round responses, *Discuss* and *Estimate* second round responses, following which judgements are combined using mathematical *Aggregation*.[17]

Prior to the elicitation the questions must be formulated, and the experts identified. Two types of questions may be asked during the elicitation: the questions of interest (for example 'Out of 1,000 born-digital files, for how many would you expect an archive to know their conditions of use?'), and the calibration questions (for example 'Out of 1,000 hard drive disks kept in a monitored commercial environment, how many drives would you expect to fail within their first 12 months of use?') with known answers which can be then used to calibrate the experts' assessments.

Pre-elicitation, the problem is defined precisely to minimize any risk of semantic or other misunderstandings arising. The data on which the calibration questions will be based is identified and finally, some training is delivered to the experts to explain what is required of them. In the first phase of the elicitation stage the experts provide individual estimates of the quantities of interest by answering the questions without discussing with, or disclosing their responses to, the other experts. They are asked to provide their estimates in a particular order: their lowest plausible, highest plausible and their best estimate of the quantities of interest. This ordering is designed to avoid anchoring the upper and lower estimates around the best estimate and leads to better accuracy.

The second phase is a facilitated discussion of the anonymized results for each question in turn, which irons out any residual semantic difficulties and allows experts to share their reasoning and any further evidence. This ensures that every expert is answering the same question based on the same evidence. The third phase is a second round of individual, private estimates, allowing experts to revise (or not) their estimates, based on what they heard in the discussion. The privacy afforded for providing second round estimates protects them from any pressure to conform to the views of others.

The calibration exercise is identical in format to the elicitation. The principal difference is that the 'answers' to the calibration questions can be checked but are not immediately accessible to the experts. The experts' estimates on the calibration questions are compared to the known values and the experts' performance can be calculated. These performance measures can be used to weight each individual's answers when they are finally combined to form a single overall estimate.

The final stage is the mathematical aggregation of experts' judgements. Commonly, some form of weighting is used based on the calibration exercise, which provides insight into the ability of the experts to estimate probabilities — a task known to be difficult. The ideal expert is both domain-savvy and good at estimating probabilities.

The elicitation workshop

In building the DiAGRAM tool, there were various areas where data was either not available or was too sparse or uncertain to be reliable. In these instances, the IDEA protocol was used to elicit quantities from our panel of experts in digital archiving. A series of workshops was held which iteratively and collaboratively produced a consensus on the interrelations between various subsystems and data that was available.[18] This also revealed where data was missing or insufficient to provide good estimates for the model. The experts were all working in archives and so the need for background papers was obviated — their experience is what we needed them to bring to the table. The missing data were formulated into questions with clear operational meanings and calibration data sought. The calibration questions were interspersed with the questions of interest in order to obfuscate which were which. This allowed us to release the experts after the initial training and clarification session to complete the questions individually. The results were processed overnight, and the facilitated discussion was held the next day. Graphs were plotted of the experts' estimates and presented in anonymized form as in Figure 2.

After the discussion, experts were again given time to adjust (or not) their first-round estimates. The new results were processed, and the final data produced and added to the model.

The final DiAGRAM model consists of the aggregated risk factors which threaten the digital archive, their respective probabilities and the dependencies between them, all expressed statistically and made available for querying and exploration via a graphical interface.

How DiAGRAM works

DiAGRAM asks the user a series of questions relating to their archive and its digital collections to provide data for the Bayesian model. These include (for example) the proportion of the records that are born-digital, the percentage of the digital material

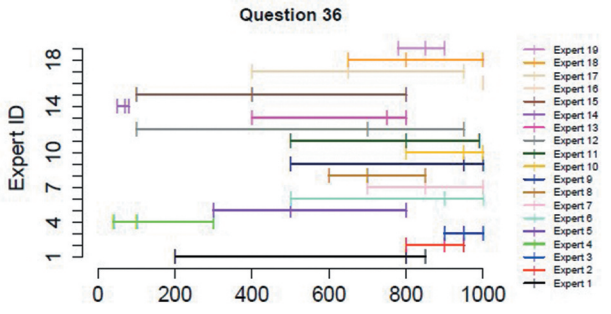

Figure 2. A range graph displaying the experts' initial estimates to question 36: 'Out of 1,000 born-digital files, for how many would you expect an archive to know their conditions of use?' Each horizontal line represents the results from a different individual. For each of these horizontal lines, the leftmost point represents the individual's estimate for the 5th percentile and the rightmost point the 95th percentile. The short vertical line in-between these end points represents the individual's estimates for the 50th percentile i.e. the median.

with a copy held offsite, and the levels achieved in assessments such as DPC RAM and the NDSA Levels of Preservation.[19]

The answers are used to adjust the data in the underlying model to reflect the user's archive and the results are expressed as two percentages: of 100 files how many of which the archive a) has intellectual control and b) can render. These two factors were deemed by the project team as the measures of success (or 'utilities') for the preservation of digital material.

These utilities are defined as:

(1) intellectual control: the probability that you have full knowledge of the file's content, provenance and conditions of use
(2) renderability: the probability that you can provide a sufficiently useful representation of the original file

Armed with the results (see Figure 3)", the user can create scenarios to explore how these probabilities would change by answering some, or all, of the questions differently and comparing the two sets of results.

Downloading these as a graph or as raw data, this evidence can be used in a business case requesting additional resources, to inform a change in policy around storage media, or to develop a preservation plan for the archive's digital collection.

Discussion

Model scope

One of the key challenges of creating DiAGRAM was deciding what to incorporate and what to leave out. The model needs to be a transferable tool that can be used by a repository as large and complex as The National Archives, as well as one that is, for example, a small charity repository run by a lone archivist. Ideally, it should be adaptable enough to be of use to para-professionals and community archives too. This meant that

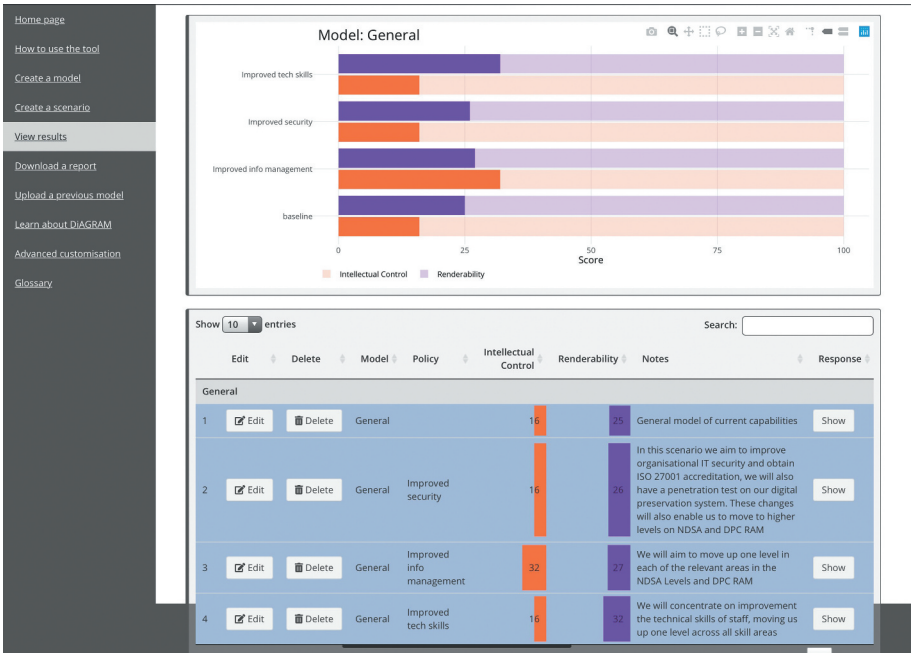

Figure 3. A snapshot of DiAGRAM's user interface. This image shows the user having set up an initial model of their archive and then creating three scenarios where they have investigated potential changes that they could make to their digital preservation practices or systems. The purple bars represent the score for renderability for the model and each scenario, and the orange the score for intellectual control.

fine-tuning the number of nodes, keeping the model to a manageable size and ensuring that the factors were readily quantifiable was key. Given the complexity of the digital landscape, this was no easy task.

The initial knowledge gathering workshop resembled a massive brainstorming exercise where archivists and technical experts drew out all of the factors that can affect our ability to effectively render and provide access to digital documents. The sequence of dependencies was then established, and this work slowly drew us towards our initial visual expression of a risk network for the digital archive. But this network had the potential to be vast and far-reaching and at this stage included such variables as target community, user type, and search facilities. This was where discussions were had, and compromises made. Some nodes were effectively amalgamated as they had no dependencies and only influenced one other node, making it justifiable to merge them together. Some nodes were collectively discounted due to an inability to quantify them or the fact that their role in the wider picture was so minimal. But there were some nodes that were more contentious. In the current iteration of DiAGRAM, the decision has been made to omit the acquisition, trust in service, and service continuity nodes that were originally part of the network (see Figures 4 and 5). It is not a decision that was reached easily and we believe that it is important that some representation of the discussions around their exclusion in particular should be provided as they remain important considerations and potential obstacles in the digital risk conversations.

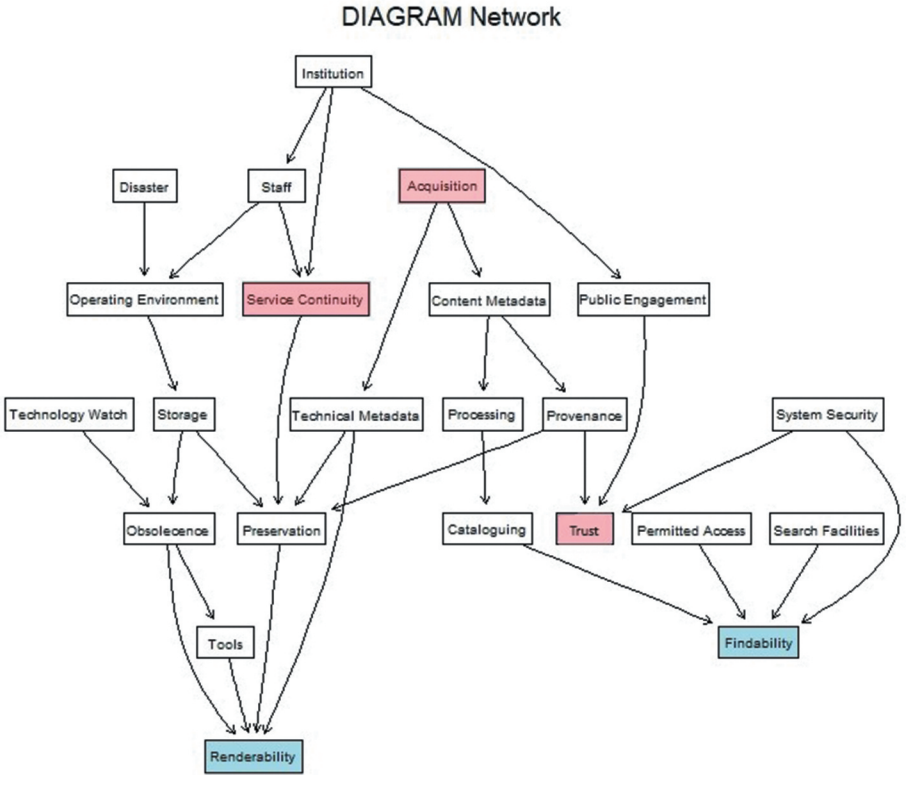

Figure 4. The network of digital preservation risks as of 31 December 2019, before acquisition, service continuity and trust were removed.

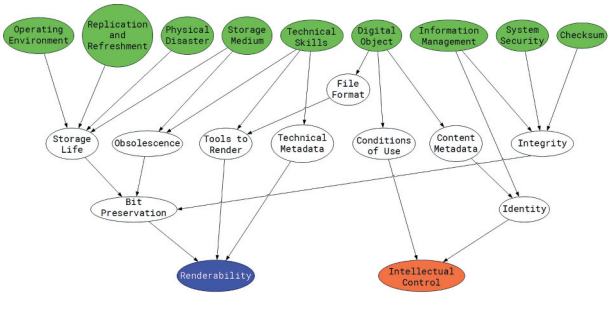

Figure 5. The latest network of digital preservation risks as of 14 October 2020.

Acquisition

Whilst it is true that preservation cannot be undertaken on a record unless it has been acquired in the first place, we agreed that the starting principle of this model would be the assumption that this has already happened and that there is material to preserve. Many factors contribute to the success (or otherwise) of the acquisition process and it could easily be an endpoint node in a larger network of its own, hence the decision not to

include it in DiAGRAM. Further, as the preservation risks under consideration are expressed from the perspective of the archive, they can only be applied to records the archive holds. Widening the model to include a consideration of whether the correct material is being acquired in their complete form crosses this conceptual boundary.

Trust in service

Initially this node was simply *trust*. Archivists and repository custodians are constantly seeking to preserve and demonstrate authenticity and strive to be a 'trusted' repository and therefore it should be of little surprise that trust began life in DiAGRAM as a key outcome node. But the discussion around what we meant by trust revealed a complex landscape of factors contributing to trust, why we wanted trust, and what we wanted trust in.

It became clear that we needed to separate out trust in the service or repository and trust in the authenticity of the digital record, as it is perfectly possible to trust the service but not the record. Once we had done this, we were able to ask: *what does trusting a digital document mean, what needs to be in place and what must we demonstrate for a user to have trust in the document?* The events which influence this were identified as a combination of measurable factors such as fixity and system security, and so trust in the digital record became the 'integrity' node and remains in DiAGRAM.

Trust in service however is more subjective and very hard to measure — many services have attempted to poll users regarding user experience and the results are always lacking in insight. Further, as we had already illustrated that it is not necessary for a user to trust the service for them to have trust in the digital records, *trust in service* was removed from DiAGRAM. Its fundamental importance may be discussed in a future paper.

Service continuity

This is arguably the most contentious exclusion because it can have a fundamental biggest impact on the feasibility and success of a digital preservation programme. And therein lies the reason it has ultimately been excluded from DiAGRAM. Archivists and collections custodians are frequently engaged in justifying their existence, proving their worth to the wider organization and the public, and advocating for more resource. They must constantly fight their corner in order to secure service continuity. For this reason, an archivist will always, rightly, argue that the key to delivering the best service they can and mobilizing the potential of the collections is stability and resource. Consequently, mention of this need arises at every opportunity.

We are by no means challenging this view, nor do we claim that digital preservation risk is immune from, and stands apart from, the need for service continuity. In fact, we have excluded service continuity as a node from DiAGRAM because we recognize its fundamental importance to the success of anything an archives service is tasked with delivering. Digital preservation is quite simply a non-starter without service continuity and the resources implicit within it, even if it is something as seemingly straightforward as having the ability to install a particular piece of software. Each node within DiAGRAM depends on resource, so situating the *service continuity* node outside of the network underlines its importance. No effective programme of work to reduce risk to the archive can take place without it. On a more granular level, most of the more measurable risks which might arise from a lack of service continuity (such as loss of data or poor knowledge

management) are already expressed elsewhere in DiAGRAM and so those particular scenarios can be modelled by examining the impacts on those nodes.

The collaborative process

The project team members came from three disciplines which do not have existing mutual connections or a history of cross-collaboration. Within the digital archiving sphere, there is a long-established tradition of collaboration (digital preservation is very much a 'team sport') and collaboration between archivists and computer scientists is becoming increasingly common. However, it is unusual for archivists to work with statisticians. Here we describe the experiences of some of the specialists on the project, each reflecting on what they learnt from the collaboration and the ways in which they adjusted their approaches to meet the needs of the other disciplines and build a strong partnership.

Experience of mathematical scientists working with archivists

There are a number of typical responses of problem owners to engaging with the mathematical sciences. One is that they believe it cannot help and are very sceptical. Another is that it is like magic and can do everything. The truth lies somewhere in between — in most problems with a quantitative aspect, mathematical sciences can provide additional insight, but can only draw out information and not create it out of thin air.

Working with The National Archives and their partner archivists has been a real pleasure. We have learned so much about archives and archiving which gives us a much richer appreciation of the value of the sector and the challenges it faces. Two of the National Archives project team have a mathematics background, which made our lives a lot easier. They were able to provide translation services when language of discipline culture threatened to obfuscate important matters. In modelling, we try to abstract the main drivers of a problem and use them to assess the effect of interventions. One challenge in dealing with enthusiastic archivists is to convey that, however interesting it is, some of the esoteric details are irrelevant for the model — we are not trying to replicate it in every aspect. It is also difficult — and this is true of many sectors — that conveying an accurate understanding of the level of precision a model can offer is difficult. Models take quantitative inputs which have variability and uncertainty associated with them, so the outputs of the models — the 'answers' cannot be more accurate than the information received. It is important to have a sense of the least significant difference between two numbers, whatever we are measuring.

It was delightful to meet so many enthusiastic archivists during the three workshops. Many expressed their fear of or aversion to maths, as society conditions us to, but still came with an open curiosity about what it could offer in the digital preservation space. They all contributed very meaningfully to the discussion on the relationships between various elements of the digital preservation system and what drives change or risk in that context. This was essential for modelling the system.

Perspective from the tool developers

The requirements for DiAGRAM were constantly changing which meant that it was important to maintain a feedback loop between the archivists and developers. In the

initial stages, simple designs were made that demonstrated the functionality of DiAGRAM. These designs were shared with The National Archives who would provide feedback that would be used in future designs. This feedback loop is what allowed us to ensure that DiAGRAM was able to evolve quickly to meet the latest requirements. Mid-development, a workshop was held that allowed us to gain more detailed feedback. This was a pivotal point in development and had a big influence on the current state of DiAGRAM. Here we were able to test different ideas as well as listen to the concerns of the archivists. It quickly became apparent that the archivists had different levels of statistical knowledge which needed to be reflected in DiAGRAM. This is what led to the development of the advanced and simple customization tabs, designed to cater for the different levels of skill within archives. We believe that these group workshops between the archivists and developers are a vital part of the requirement gathering process and allow us to include more people in the feedback loop, ultimately making DiAGRAM a better platform for archivists.

The main development motive was to produce an interface that is the most intuitive for an archivist, however, coming from a computer science background, our interpretation of the requirements was often different. A feature that seemed to us to be easy to use, user input for the conditional probabilities of the nodes, was reported to be confusing and ambiguous by the beta version users. We initially decided to develop it as a numeric input, but the archivists wanted a more familiar experience, such as slider input or radio buttons. It was all about tuning our thought process before being able to have effective inter-disciplinary collaboration.

Another unique factor during this collaboration was the maintenance of DiAGRAM. In a typical computer science setting, the post-development changes are handled by developers with similar skills to those that wrote the application. Due to the nature of the project and our limited availability however, the changes were managed by the Project's Research Assistant, who has a less advanced knowledge of the development environment and tools. Due to this, we employed coding techniques that would allow anyone with basic coding knowledge to make changes. For example, instead of embedding the user input fields and node labels within the source code, we created a text file and made the code read from that file, to make the tool more readily configurable. Any change to the corresponding field in the text file is reflected in the user interface. This will make the tool more easily sustainable for the future.

Working with the archivists from The National Archives and others such as Transport for London has broadened our perspective around the translation of project requirements to source code and understanding the desired user experience.

Archivists' perspective

The first workshop introduced participants to Bayesian network principles and theory. Most of the archivists had never been exposed to Bayesian networks prior to this project. Following some initial confusion, working through a number of exercises helped the principles and methods become clear. The presence of partners with different areas of expertise in the room enabled further learning exchanges between participants during the workshops as archival workflows and processes were discussed to decide upon nodes. As each node was examined and debated between groups, it was quite valuable to hear challenges to my own

domain and assumptions in the round. Due to the range and depth of expertise within the group this ensured that elements outside of each individual's expertise were taken into consideration.

The advantage of creating DiAGRAM in this manner meant that the discussion about each variable took into account technical, archival and statistical factors. As the work progressed and the tool developed, we could see how assumptions and perceptions of risks within our various organizations could be adapted to the model. The flexibility of the tool enables an archive of any size to analyse collections on both a macro and micro level. Risks can also be analysed as conditions change within the archive over time. Outputs from the tool can be used when making business decisions regarding resource allocation and effort for the life cycle of digital content and collections.

I was initially concerned that my relatively traditional archival background would not have furnished me with a strong enough grounding in probability and statistics. To some extent this proved to be true. There are details of how the model works that I cannot explain, and I have relied on more qualified team members to understand and make the right decisions. I would therefore, not only agree with the Statisticians' perspective on how much the inclusion of two mathematicians in The National Archives' team aided building the model, I would say they were a large part of its success. They were able to help translate between statistical and archival frames of reference and this proved hugely useful, especially in the early stages of the project when the partners from the two disciplines were learning about each other's subject areas.

Working with mathematicians on this project has demonstrated just how much a discipline that we thought completely unrelated to archives can help the profession tackle problems that were previously considered intractable. All that was needed was the willingness to engage with unfamiliar subjects, to ask lots of questions and relinquish some assumptions (and fears). As archivists we had the opportunity to talk about our profession to a fresh audience which was not only interested but wanted to understand it from a different point of view which made us examine our work through a different lens.

Conclusions and further work

Extensions of research

During the creation of DIAGRAM, we encountered several areas where the digital archives sector as a whole lacks quantitative data. Although expert elicitation protocols such as the *IDEA* protocol described above can be effective in filling these gaps, there is no question that effective management of digital preservation risk requires a sound evidence base to inform decisions. Some of the gaps we encountered will always benefit from elicitation approaches whilst others would be better addressed through more structured approaches to data collection in digital archives. This may be complex and, in some areas, coordinated projects will be required to examine and understand requirements and build the evidence base that is needed. However, there are also a number of simple steps that archives can

take today which would vastly improve the accuracy and availability of good data to support the long-term preservation of digital material, now and far into the future.

We recommend an increased focus on the collection of operational data in digital archives. This data may arise from direct actions taken by archivists to manage the material themselves, and from actions that relate to the wider infrastructure and policy environment in which the digital archive operates.

For example, improved self-monitoring of storage media failures will provide quantitative data that will help put this risk into context. Existing approaches to data capture should be extended to record more comprehensive and structured information on the media itself: not only make, batch, capacity and date of creation, which many archives already record, but factors such as how long the media was in operation before it was replaced, the reason for replacement, how often it had been read and written to and how much of its capacity was used. It is important to recognize the value of such data not only for resource planning, but as part of the preservation process itself.

Similarly, we recommend increased internal recording of preservation actions such as fixity checks and, most importantly, their frequency and outcome. Even archives that take analogue only accessions can assist by defining the level of metadata they consider to be sufficient for their organization and recording the extent to which deposited material meets this standard.

At the time of writing, we are in the midst of a global pandemic caused by COVID-19 and many countries are in a national lockdown. Undoubtedly, there will be research that will follow in the next few years looking at the effect this has had on the archives sector and we also hope that there will be a new pool of evidence available, from which we can evaluate the real impacts of this severe and rare event on digital preservation risks. This will give us the opportunity to re-evaluate the best way to incorporate service continuity and sustainability threats into DiAGRAM.

Future of DiAGRAM

We want DiAGRAM to be a tool that is of continuing value to the archives sector so feedback at this early stage is critical. Our long-term ambition is that DiAGRAM will be developed collaboratively by the archival community and managed by The National Archives on the same model as existing digital preservation tools such as DROID (Digital Record Object Identification) and the technical registry PRONOM.[20] We envisage DiAGRAM as a living service, which we will regularly review and update with new evidence as it becomes available and as the challenges of preserving the Nation's digital heritage continue to emerge and change.

Notes

1. The National Archives, "Digital Strategy," 2.
2. The National Archives and Pye Tait Consulting, "Archives Workforce Development Strategy", 14-15; and The National Archives, "Digital Capacity Building Strategy".
3. If nothing else, storage hardware is typically replaced on a more frequent cycle than this, though this may be invisible to the end user.
4. Conway, *Preservation in the Digital World*, Context For Action.
5. HM Government, *The Orange Book*.
6. International Organization for Standardization, *ISO 31000:2018 Risk Management — Guidelines*; Barateiro, Antunes, and Borbinha, "Proposals for New Perspectives"; and Barateiro et al., "Designing Digital Preservation Solutions".
7. Vermaaten, Lavoie, and Caplan, "Threats to Digital Preservation".
8. Ibid., Appendix.
9. Rosenthal et al., "Requirements for Digital Preservation Systems"; Consultative Committee for Space Data Systems, "Open Archival Information System"; and see http://www.iso16363.org/standards/.
10. National Digital Stewardship Alliance, "Levels of Digital Preservation".
11. The National Archives, "Archive Service Accreditation Guidance," 63.
12. The National Archives, "Digital Strategy".
13. Fenton and Neil, *Risk Assessment with Bayesian Networks*.
14. Rafiq, Chryssanthopoulos and Sathananthan, "Bridge Condition Modelling".
15. For more information on the Turing Institute see https://www.turing.ac.uk/; and for more information on the Applied Statistics and Risk Unit see https://warwick.ac.uk/fac/sci/statistics/asru/.
16. European Food Safety Authority, "Guidance on Expert Knowledge Elicitation". This publication gives further information on how to use structured expert judgement elicitation.
17. Hanea et al., "Investigate Discuss Estimate Aggregate".
18. Barons, Wright and Smith, "Eliciting Probabilistic Judgements".
19. The existing assessments referenced in DiAGRAM are the Digital Preservation Coalition, "DPC Rapid Assessment Model"; National Digital Stewardship Alliance, "Levels of Digital Preservation"; and DigiCurV, "DigiCurV Curriculum Framework".
20. For more information about digital preservation tools, see https://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/preserving-digital-collections/digital-preservation-tools-systems/.

Funding

This work was supported by the National Lottery Heritage Fund under project reference number OM-19-01060; The Engineering and Physical Sciences Research Council under grant EP/R511808/1; and The National Archives (UK)

Notes on contributors

Dr Martine J. Barons is the director of the Applied Statistics & Risk Unit at the University of Warwick. Martine has research interests in all aspects of decision support and was part of the team, led by Jim Q. Smith, which developed the IDSS paradigm. Martine has applied decision support paradigms, most notably for pollinator abundance and household food security. She also has extensive experience in structured expert judgment, both within the above applications and as a consultant for government applications.

Sidhant Bhatia is a Software Engineer Graduate from Monash University. During the course of the project, he was undertaking a short-term research placement at the University of Warwick under the supervision of Dr Martine J. Barons. He joined the development team to build the Graphical User Interface dashboard for the statistical model and was primarily involved in the design process to ensure an intuitive user experience. Prior to the placement, he had industrial experience in developing web applications, both front-end and back-end. He also explored and developed basic statistical models as a personal interest.

Jodie Double is the Digital Content and Copyright Manager at Leeds University Library. She has over two decades of experience working in digital collections and archives and joined the University of Leeds in 2009 from the University of Minnesota as Director of Digital Collections and Archives in the College of Design. Her role at Leeds for the past 11 years has focused on developing access and services around the growing corpus of digital content created from the extensive collections at Leeds in addition to the growing body of born-digital content being deposited.

Dr Thaís C. O. Fonseca is an Associate Professor at UFRJ, Brazil and a Research Fellow at the Applied Statistics & Risk Unit at the University of Warwick. Thaís' main research interests include Bayesian inference for time series and stochastic processes, Bayesian network models, and Bayesian econometrics. She has experience in consultancy for industry and government with applications to insurance and economics.

Alex Green is the Service Owner for Digital Preservation at The National Archives and is the project lead. She is an experienced Digital Archivist having worked on creating user-centric digital tools and services for the past twenty years.

Stephen Krol is a Computer Scientist/Data Scientist graduate from Monash University, Melbourne, Australia. He currently works as a data science consultant and is experienced in Python and R. Through a partnership between Warwick University and Monash University he was able to build the model for this project.

Hannah Merwood is a Research Assistant in Applied Statistics at The National Archives (UK) on secondment from the Department for Digital, Culture, Media and Sport. She holds a bachelor's degree in mathematics and statistics from the University of Oxford and is a member of the Government Operational Research Service analytical profession. Hannah has experience of developing complex models and data tools to support decision makers within government, including for prisoner escort contracts and broadband delivery programmes.

Dr Alec Mulinder is Head of Service Assurance at The National Archives (UK). He is responsible for making sure that new digital services, and enhancements to existing digital services, meet the required quality standards set by UK government. Alec has been involved in the DiAGRAM project from the beginning, researching the applicability of Bayesian modelling techniques, and being a main contributor to our successful National Heritage Lottery Fund bid. His interest in digital preservation risk began in 2012 working on the UK Government funded Digital Continuity project. Alec is also working on modelling of digital preservation storage, is a co-supervisor of two PhD students researching born digital access and has a PhD in Medieval History.

Dr Sonia Ranade is Head of Digital Archiving at The National Archives (UK), with responsibility for digital services to depositors (for selection and transfer), preservation of the digital public record and access to digital records. Her research interests include digital preservation risk, probabilistic approaches to archival description and developing new access routes for digital archives. Sonia holds a PhD in Information Science.

Prof Jim Q. Smith is a Professor of Statistics at Warwick University and a Fellow of the Alan Turing Institute. He is a decision analyst and Bayesian dynamic systems modeller with interests that span statistical inference, data science, machine learning and operations research. He specialises in the methodology and application of various types of graphs for describing uncertain processes and also

in expert elicitation, especially structural elicitation. He has recently worked with domain experts to design Bayesian decision support systems for managing risks associated with nuclear accidents, public health, policing, food poverty and COVID 19. He has published over 200 refereed papers and written three books.

Tamara Thornhill is the Corporate Archivist at Transport for London (TfL) and has 16 years' experience working in archives and records management roles. She has been at TfL since 2010, responsible for the management of a collection consisting of over 165,000 files of physical material dating from the 16th century, and 20TB of digital files. She firmly believes that archives are a real asset to their parent body and wider communities and that they should be accessible and used. Since joining TfL, Tamara has undertaken various outreach activities to promote the Archives service. This programme has seen overall enquiries rise by 125% and visits increase by 359%. Exhibitions have expanded, and exhibition attendance has increased by 802%.

David Underdown joined The National Archives in 2005 as a database administrator and soon gained his introduction to digital preservation from supporting our PRONOM registry of file formats and involvement in projects to refine and update our digital repository system. Since David's background (a degree in mathematics from Imperial College London and several years working in systems development for a life and pensions company) had not really prepared him for working in archives, he used a general interest in First World War history to develop his experience of archival research and archival theory. David is also involved in defining image and metadata specifications for digitisation projects such as First World War Unit Diaries and 1921 Census. His current research project sees a return to his mathematical roots, applying Dynamic Bayesian Networks to modelling digital preservation risk through the National Heritage Lottery fund supported project 'Safeguarding the Nation's Digital Memory'.

ORCID

Martine Barons http://orcid.org/0000-0003-1483-2943
Sidhant Bhatia http://orcid.org/0000-0003-1650-0520
Jodie Double http://orcid.org/0000-0002-7152-3947
Thais Fonseca http://orcid.org/0000-0002-4943-3259
Alex Green http://orcid.org/0000-0003-2463-3649
Stephen Krol http://orcid.org/0000-0002-9474-3838
Hannah Merwood http://orcid.org/0000-0002-0520-9319
Alec Mulinder http://orcid.org/0000-0002-8900-3798
Sonia Ranade http://orcid.org/0000-0002-2674-8370
Jim Q Smith http://orcid.org/0000-0002-9224-5317
Tamara Thornhill http://orcid.org/0000-0001-9287-2399
David H Underdown http://orcid.org/0000-0002-8123-4655

Bibliography

Barateiro, José, Gonçalo Antunes, and José Borbinha. "Addressing Digital Preservation: Proposals for New Perspectives." 2009.

Barateiro, José, Gonçalo Antunes, Filipe Freitas, and José Borbinha. "Designing Digital Preservation Solutions: A Risk Management-Based Approach." *International Journal of Digital Curation* 5, no. 1 July 21 (2010): 4–17. doi:10.2218/ijdc.v5i1.140.

Barons, Martine J., Sophia K. Wright, and Jim Q. Smith. "Eliciting Probabilistic Judgements for Integrating Decision Support Systems." In *Elicitation: The Science and Art of Structuring Judgement*, edited by Luis C. Dias, Alec Morton, and John Quigley, 445–478. Cham: Springer International Publishing, 2018. International Series in Operations Research & Management Science. doi:10.1007/978-3-319-65052-4_17.

Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS)." no. 2 (2012): 135.

Conway, Paul. *Preservation in the Digital World*. Council on Library and Information Resources, 1996. https://www.clir.org/pubs/reports/conway2/index/ .

"CoreTrustSeal." *CoreTrustSeal*. Accessed May 1, 2020. https://www.coretrustseal.org/ .

DigiCurV. "DigiCurV Curriculum Framework." Accessed October 12, 2020. https://www.digcurv.gla.ac.uk/skills.html .

Digital Preservation Coalition. "Digital Preservation Coalition Rapid Assessment Model." September, 2019. doi: 10.7207/dpcram19-01

Dobratz, S., and A. Schoger. "Trustworthy Digital Long-Term Repositories: The Nestor Approach in the Context of International Developments." *Research and Advanced Technology for Digital Libraries* 4675 (2007): 210–222. ECDL 2007. Lecture Notes in Computer Science.

"DRAMBORA: Digital Repository Audit Method Based on Risk Assessment." Accessed May 1, 2020. http://www.repositoryaudit.eu/ .

European Food Safety Authority. "Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment." *EFSA Journal* 12, no. 6 (2014): 3734. doi:10.2903/j.efsa.2014.3734.

Fenton, N., and M. Neil. *Risk Assessment and Decision Analysis with Bayesian Networks*. 2nd ed. New York: Chapman and Hall/CRC, 2019. doi:10.1201/b21982.

Hanea, A. M., M. F. McBride, M. A. Burgman, B. C. Wintle, F. Fidler, L. Flander, C. R. Twardy, B. Manning, and S. Mascaro. "Investigate Discuss Estimate Aggregate for Structured Expert Judgement." *International Journal of Forecasting* 33, no. 1 January 1 (2017): 267–279. doi:10.1016/j.ijforecast.2016.02.008.

HM Government. *The Orange Book: Management of Risk – Principles and Concepts*. Accessed May 1, 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/866117/6.6266_HMT_Orange_Book_Update_v6_WEB.PDF .

International Organization for Standardization. *ISO 16363:2012 Space Data and Information Transfer Systems — Audit and Certification of Trustworthy Digital Repositories*. 1st ed. 2012. Geneva: International Organization for Standardization https://www.iso.org/obp/ui/#iso:std:iso:16363:ed-1:v1:en .

International Organization for Standardization. *ISO 31000:2018 Risk Management — Guidelines*. 2nd ed. 2018. Geneva: International Organization for Standardization https://www.iso.org/obp/ui#iso:std:iso:31000:ed-2:v1:en .

The National Archives. 2017. "Digital Strategy." Accessed July 15, 2020. https://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf .

The National Archives. 2018. "Archive Service Accreditation: Guidance for Developing and Completing an Application." Accessed July 17, 2020. https://www.nationalarchives.gov.uk/documents/archives/archive-service-accreditation-guidance-june-2018.pdf .

The National Archives. "Plugged In, Powered Up: A Digital Capacity Building Strategy for Archives." Accessed May 1, 2020. https://www.nationalarchives.gov.uk/documents/archives/digital-capacity-building-strategy.pdf .

The National Archives. "Pronom." Accessed June 29, 2020. https://www.nationalarchives.gov.uk/PRONOM/Default.aspx .

The National Archives. "File Profiling Tool (DROID)." Accessed June 29, 2020. https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/ .

The National Archives, and Pye Tait Consulting. "Archives Sector Workforce Development Strategy." August, 2018. https://www.nationalarchives.gov.uk/documents/archive-sector-workforce-strategy.pdf .

National Digital Stewardship Alliance. "Levels of Digital Preservation." *National Digital Stewardship Alliance - Digital Library Federation*. Accessed June, 29 2020. http://ndsa.org//publications/levels-of-digital-preservation/ .

Rafiq, M. Imran, Marios K. Chryssanthopoulos, and Saenthan Sathananthan. "Bridge Condition Modelling and Prediction Using Dynamic Bayesian Belief Networks." *Structure and Infrastructure Engineering* 11, no. 1 January 2 (2015): 38–50. doi:10.1080/15732479.2013.879319.

Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito. "Requirements for Digital Preservation Systems: A Bottom-Up Approach." *D-Lib Magazine* 11, no. 11 (November, 2005). doi:10.1045/november2005-rosenthal.

Vermaaten, Sally, Brian Lavoie, and Priscilla Caplan. "Identifying Threats to Successful Digital Preservation: The SPOT Model for Risk Assessment." *D-Lib Magazine* 18, no. 9/10 (September, 2012). doi:10.1045/september2012-vermaaten.