Safeguarding the Nation's Digital Memory: Bayesian Network Modelling of Digital Preservation Risks

Martine J. Barons, Thais C. O. Fonseca, Hannah Merwood, and David H. Underdown

Abstract Archives comprise primary sources which may be physical, born digital or digitised. Digital records have a limited lifespan, through carrier degradation, software and hardware obsolescence and storage frailties. It is important that the original bitstream of these primary sources is preserved and can be demonstrated to have been preserved. Soft elicitation with experienced archivists was used to identify the most likely elements contributing to digital preservation success and failure and the relationships between these elements. A Bayesian Network representation of an integrating decision support system provided a compact representation of reality, enabling the risk scores for various scenarios to be compared using a linear utility function. Thus, the effect on risk of various actions and interventions can be quantified. This tool, DiAGRAM, is now in use.

1 Introduction

Archives comprise primary sources which can be physical, born digital and digitised. Digital records have a limited lifespan, through carrier degradation, software and hardware obsolescence and storage frailties. It is important that the original bitstream of these primary sources is preserved and can be demonstrated to

M. J. Barons (✉)
AS&RU, Department of Statistics, University of Warwick, Coventry, UK
e-mail: martine.barons@warwick.ac.uk

T. C. O. Fonseca
Department of Statistical Methods, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: thais@im.ufrj.br

H. Merwood
Government Operational Research Service, London, UK

D. H. Underdown
The National Archives, Surrey, UK
e-mail: david.underdown@nationalarchives.gov.uk

M. Ehrhardt, M. Günther (eds.), *Progress in Industrial Mathematics at ECMI 2021*,
The European Consortium for Mathematics in Industry 39,
https://doi.org/10.1007/978-3-031-11818-0_65

have been preserved; this consumes significant resources [1]. Digital preservation (DP) is crucial for ensuring the longevity of societal history, for research, legal accountability, government and business planning. It is a maturing field, with the main standards around 20 years old. Larger (and relatively better funded) archives such as national, state and provincial archives in high income countries have been engaged in DP for longer periods. It is now becoming a pressing issue for all archives.

The archival sector typically lacks sufficient people, sufficiently skilled people and sufficient funding to undertake all possible mitigations against these risks. Thus, there is a need for support in choosing the mitigation strategies which bring the largest and most immediate reduction in overall risk levels in the current context of an individual archive: there is not a 'one-size fits all' solution.

The National Archives in the United Kingdom (TNA), and the Applied Statistics & Risk Unit (AS&RU) at the University of Warwick collaborated to build decision support suitable for identifying risks to digital archives and quantifying the efficacy of mitigation strategies, the Digital Archiving Graphical Risk Assessment Model (DiAGRAM) [2].

2 Methodology

Soft elicitation [3] with experienced archivists was used to identify the most likely elements contributing to digital preservation success and failure and the relationships between these elements. This established, it became obvious that a Bayesian Network [4, 5] representation of an integrating decision support system (IDSS, [6, 7] would be appropriate as a compact representation of reality in this case. However, not all the data required to quantify the model was available, so structured expert judgement was employed to provide data in the gaps. The IDSS is a new paradigm for drawing together evidence from different parts of large systems to provide decision support. Each part of the system is typically overseen by a panel of domain experts using their own data and, often complex, models. Panels contribute key summaries of future expectations under different candidate policy decisions. The IDSS then allows the decision centre to calculate expected utility scores for these candidate policies for comparison and decision support.

2.1 Bayesian Networks

A discrete Bayesian Network as defined in [4] is a compact representation of the joint probability distribution $p(\mathbf{x})$ of a p-variate vector of random variables $\mathbf{X} = (X_1, \ldots, X_p)'$. The model is specified by the set $\mathcal{N} = (\mathcal{X}, \mathcal{G}, \mathcal{P})$ with elements given by

1. a graph $\mathcal{G} = (V, E)$ with nodes V and connections E;
2. a set o variables $\mathcal{X}$ representing the nodes of $\mathcal{G}$;
3. a set of conditional distributions $\mathcal{P}$ with distribution $p_i(x_i \mid x_{pa(i)})$ for each $X_i \in \mathcal{X}$,

where $X_{pa(i)}$ is the set of parents of X_i. A Bayesian Network model is composed by the representation induced by $\mathcal{N}$ which is given by

$$p(\mathbf{x}) = \prod_{v \in \mathcal{X}} p_i(x_v \mid x_{pa(v)}).$$

The inferential problem depends on the computation of $P(X_v = x_v \mid \epsilon)$, $X_v \in \mathcal{X}$ given a set of evidences ϵ, that is, the computation of total probabilities depending on sums and multiplications. However, this computation is costly even for small p. Often algorithms such as Logic Sampling are used to approximate the predictive probabilities of interest. In the context of categorical data, the distributions assumed for the observations are multinomial such that $X_i \mid X_{pa(i)} = j, \boldsymbol{\theta}_{ij} \sim Mult(M_{ij}, \boldsymbol{\theta}_{ij})$ and are represented as conditional probability tables (CPTs). If a Dirichlet prior with parameter $\boldsymbol{a}_{ij}$ is assumed for $\boldsymbol{\theta}_{ij}$ then the posterior distribution is Dirichlet with parameter $\boldsymbol{N}_{ij} + \boldsymbol{a}_{ij}$ with N_{ijk} the counts of $\{X_{ik} = x_{ik}\}$ when $\{X_{pa(i)} = j\}$. For a practical guide on how to perform inference and prediction using Bayesian Networks see [8].

Where data was not available, Structured Expert Judgement (SEJ) was used to quantify experts’ uncertainties on the values for the conditional probability tables.

2.2 *Structured Expert Judgement*

Expert judgement is pervasive in all forms of risk analysis [7]. Structured expert judgement elicitation is a well-established paradigm for eliciting expert judgements of uncertain quantities and event occurrences [9]. Structured protocols seek to mitigate the most pervasive cognitive frailties when asking for subjective judgements, such as group-think, availability bias, personality effects and overconfidence. We used the recently-developed IDEA protocol [10]. Calibration questions are included, drawn from existing surveys and reports, on which individual experts’ accuracy and informativeness can be calculated for performance-weighted pooling of the results into a single distribution, using the classical approach [11].

3 DiAGRAM

The Digital Archiving Graphical Risk Assessment Model (DiAGRAM) is a bespoke tool developed to facilitate the computation of digital preservation risks and provide comparison of competing policies. It aims to improve users’ understanding of digital

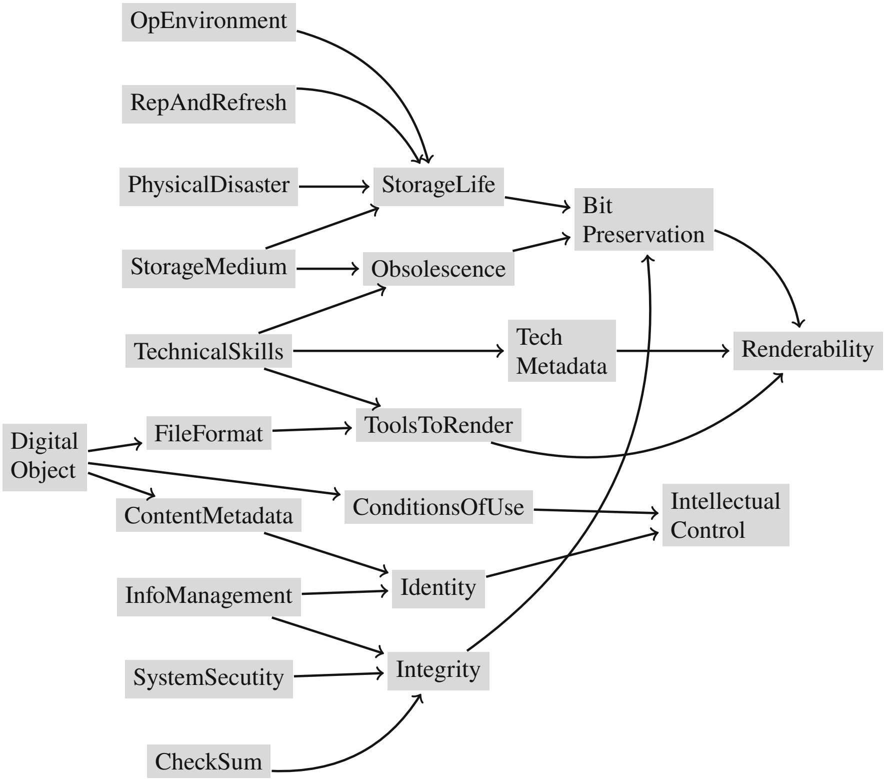


Fig. 1 Qualitative description of the digital preservation system

archiving risks, empower archivists to compare and prioritise different threats to the digital objects and to aid in quantifying the impact of risk events and risk management strategies on digital preservation to support decision making.

The model contains the network elicited using soft elicitation $\mathcal{G}$ and the conditional probability tables $\mathcal{P}$ representing the uncertainty in the nodes obtained via historical data when it is available, and through SEJ elicitation otherwise.

3.1 Network Structure Construction

The variables and the qualitative relationships between them were elicited through close communication with domain experts. The experts' collective views were represented by a Directed Acyclic Graph (DAG) (Fig. 1). The variables included in the model were Digital Object, Identity, Conditions of Use, Intellectual Control, Information Management, Technical Skills, Operating Environment, Content Metadata, Technical Metadata, Checksum, File Format, Bit-preservation, Obsolescence, Tools to render, Storage Medium, Storage Life, Replication and Refreshment, System Security, Integrity and Renderability.[1] The variables Intellectual Control and Renderability comprise the utility function which provides comparative scores for candidate policies in the policy comparison step. See [2] for further details on structure construction and node definitions.

3.2 *Expert Elicitation Results*

SEJ was used in DiAGRAM to quantify Storage life, Obsolescence, Technical Metadata, Tools to Render, Conditions of Use, Content Metadata, Identity, Integrity, Bit Preservation and Renderability. In the elicitation sessions, 22 participating experts answered 20 calibration questions and 24 questions of interest. The transformed Kullback-Leibler divergence and the performance-weighted outcomes for all experts are presented in Table 1. The results show experts 8, 12 and 16 had the best performances on the calibration questions and experts 13, 20, and 21 had the worst performances.

3.3 *Joint Probability Distribution*

This section computes the probability distributions based on the structure, tables elicited from experts and data available. The data sources used were: the 2019 JISC digital skills survey of over 300 UK archive professionals; the cloud data storage providers on access and durability; data from the Environment Agency on the long-term flood risk of UK postcodes; and data from TNA on file formats by digital object type.

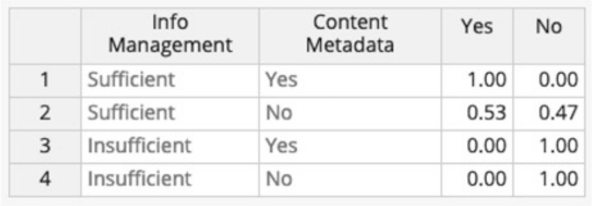
In DiAGRAM, of the 21 nodes, 9 have the probabilities customisable by the users to reflect their institution: Digital Object, Operating Environment, Replication and Refreshment Storage Medium, Technical Skills, Information Management, System Security and Checksum.

For comparative purposes DiAGRAM provides a Baseline Model (BM) where the customisable nodes are set to: (1) no technical skills; (2) good level of system security (74%); (3) 0% of files have a check-sum; (4) 14% of files have sufficient internal information management systems in place; (5) 100% of the digital archive is born digital; (6) 100% of storage media are stored on outsourced (cloud) storage; (7) 100% of files have a good replication and refreshment strategy in place; (8) operating environment was considered 100 % as all files have copies in different locations; (9) The risk of physical disaster (flood risk rating) is very low. For this baseline model, the conditional probability table for the node Identity obtained in

[1] See DiAGRAM's 'Glossary' tab here: https://nationalarchives.shinyapps.io/DiAGRAM/.

Table 1 Experts ID, experts transformed Kullback-Leibler (KL) divergence and final combined weights ($\times 10^3$). Weights close to 0 indicate worse performances and close to 1 ($\times 10^3$) indicate better performances

Expert ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
KL divergence	14	11	10	25	20	9	36	6	37	25	17	6	45	19	19	6	24	26	28	48	66	8
Final weight	7	25	33	0	0	38	0	229	0	0	3	209	0	1	1	295	0	0	0	0	0	158



	Info Management	Content Metadata	Yes	No
1	Sufficient	Yes	1.00	0.00
2	Sufficient	No	0.53	0.47
3	Insufficient	Yes	0.00	1.00
4	Insufficient	No	0.00	1.00

Fig. 2 Conditional probability table for the node Identity obtained in DiAGRAM for the baseline Commercial Backup model

DiAGRAM for this setup is presented in Fig. 2. Probability tables for all nodes can be obtained and are used to compute the final utility function.

3.4 Utility Computation and Scenario Evaluation

In consultation with a wide range of digital archivists, the utility for DiAGRAM was defined as Renderability and Intellectual Control. Renderability (R) captures the need for the digital object to have a sufficiently useful representation of the original file. 'Sufficiently useful' depends on the use to which a digital object is being put. Intellectual Control (IC) is the archivist's need to have full knowledge of the digital object's content, provenance and conditions of use. IC requires sufficient metadata that the archivist can identify the appropriate object, see how it relates to other objects from the same source, and understand whether they have the copyright permissions to make reproductions, or if data protection, etc. prevents the object from being made publicly available (and how long those restrictions will remain applicable).

We compare the Baseline Model with the alternative scenario of Commercial Backup (CB), which is as for BM but improving information management to 43 % and technical skill level to 30 %. The risk scores for BM and the CB are compared using a linear utility function (Fig. 3). The CB scenario has larger total score (62: IC = 20, R = 42) than BM (44: IC = 6, R = 38), showing that moving to CB improves digital preservation.

Acknowledgments The authors acknowledge with gratitude valuable contributions to the project by staff from partner institutions,[2] and by the additional experts who participated in the elicitation. 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).

[2] See project webpage: https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/safeguarding-the-nations-digital-memory/.

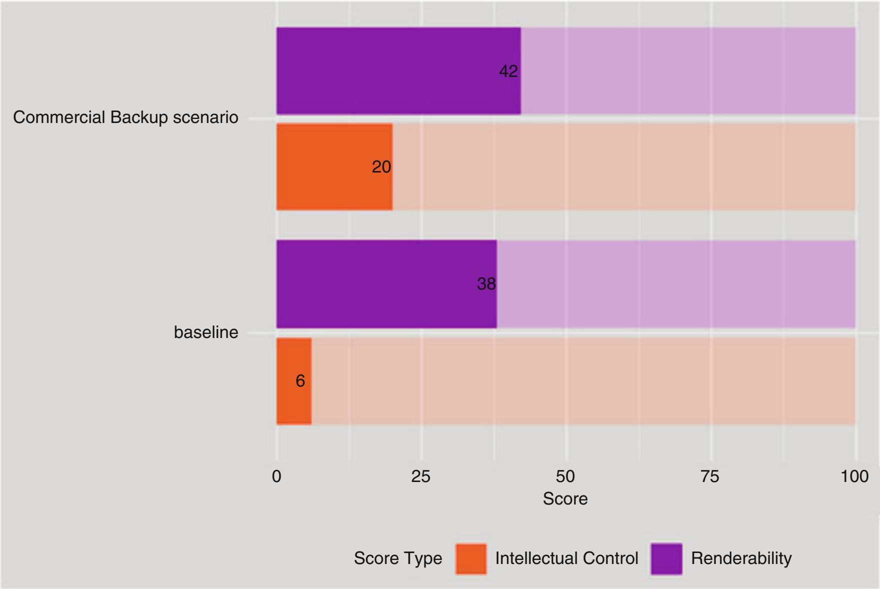


Fig. 3 Intellectual control and renderability scores comparison for the baseline commercial backup model and the commercial backup scenario

References

1. R.D. Frank. The social construction of risk in digital preservation. Journal of the Association for Information Science and Technology, 71(4):474–484, 2020.
2. M. Barons, S. Bhatia, J. Double, T. Fonseca, A. Green, S. Krol, H. Merwood, A. Mulinder, S. Ranade, J.Q. Smith, T. Thornhill, and D.H. Underdown. Safeguarding the nation's digital memory: towards a Bayesian model of digital preservation risk. Archives and Records, 42(1):58–78, 2021.
3. Simon French. From soft to hard elicitation. *Journal of the Operational Research Society*, pages 1–17, 2021.
4. F. Jensen and T.D. Nielsen. Bayesian networks and decision graphs. Springer, 2007.
5. J.Q. Smith. Bayesian decision analysis: principles and practice. Cambridge University Press, 2010.
6. J.Q. Smith, M.J. Barons, and M. Leonelli. Coherent frameworks for statistical inference serving integrating decision support systems. arXiv preprint arXiv:1507.07394, 2015.
7. M.J. Barons, S.K. Wright, and J.Q. Smith. Eliciting probabilistic judgements for integrating decision support systems. Springer, New York, New York, USA, 2018.
8. M. Scutari and J.-B. Denis. Bayesian Networks: With Examples in R. CRC Press, 2014.
9. F. Bolger, A. Hanea, A. O'Hagan, O. Mosbach-Schulz, J. Oakley, G. Rowe, and M. Wenholt. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal, 12(6):Parma, Italy, 2014.
10. A. Hanea, M. McBride, M. Burgman, B. Wintle, F. Fidler, L. Flander, S. Mascaro, and B. Manning. $\text{I}_{\text{nvestigate}}\text{D}_{\text{iscuss}}\text{E}_{\text{stimate}}\text{A}_{\text{ggregate}}$ for structured expert judgement. International Journal of Forecasting, 33(1):267–279, 2016.
11. Roger Cooke, Max Mendel, and Wim Thijs. Calibration and information in expert resolution; a classical approach. Automatica, 24:87–93, 1988.