Safeguarding the nation's digital memory

Martine J. Barons, CMath

Director
Applied Statistics & Risk Unit
University of Warwick

Martine.Barons@warwick.ac.uk

ECMI21 13th April 2021

Overview

1. Archives
   - The digital preservation problem
2. IDSS
3. The National Archives Project
   - Soft Elicitation
   - Data
   - SEJ
   - The Tool
4. Lessons
   - Lockdown
   - Digital Preservation Awards

What are archives?

Collections of information known as records

Records

- letters
- reports
- minutes
- registers
- maps
- photographs and films
- digital files
- sound recordings

Records in an archive are primary sources. Archives provide first-hand information or evidence relating to historical events or figures.

Archives are managed by a variety of types of institutions and the materials they collect differ.

Archives

- Government or national archives: materials related to all levels of government
- Corporate archives: manage and preserve business records.
- College and university archives:preserve materials related to the institution.
- Historical societies: preserve materials related to a specific region, event, or industry.
- Museum archives: diverse collections typically consisting of artwork or artifacts
- Religious archives: collect and preserve materials related to a faith, denomination or place of worship
- Special collections: a collection of items that are either irreplaceable or rare, usually in a library.

The National Archives

TNA

- Official archive of the UK Government
- Preserve the key records created by around 250 government departments e.g. Multiple email threads on a decision around public safety which will have rich embedded metadata, multiple attachments and image footers; AI algorithms used to decide who has the right to remain in the country
- At last count the digital archive held 4759.5TB of records
- Records are text, videos, sound recordings, databases, 3d models, images etc.
- Have to preserve those records forever
- Socially important records e.g. London Olympics
- Legally important records e.g.Hillsborough 15 April 1989

Digital Preservation

The act of preserving digital records is termed digital preservation. There is a wide variety of risks to digital files.The risks are connected and influence one another.

- Digital preservation is a relatively new concern, especially for archivists, many of whom are still working primarily with analogue materials.
- It takes time to get to a point where professionals know what data they need to start monitoring in order to better make decisions.
- Sharing data on risks means admitting something has gone wrong, which understandably might be something many organisations and individuals will be cautious to do.
- Obsolescence and limited lifespan: Archivists like to say that digital records last for ever - or 5 years, whichever comes first.

Digital Preservation challenges

Need to preserve the original bitstream and be able to render it 'sufficiently' for use.

Challenges

- Software - wordstar, bespoke software - emulators
- Storage medium - Facebook, Myspace, Google (reduced volume), cloud, account access
- Storage hardware - floppy disc, 8-track, VHS / Betamax, phone
- Storage life - laptop, flash drive, hard drive, SD card, magnetic tape, malware corruption
- Copying errors
- Storage compression - photographs on Facebook or Google
- Natural disaster or accident - Fire, flood, earthquake - location

Integrating Decision Support systems

A formal & defensible statistical methodology to draw together inferences when:

- Users are decision *Centres*
- Centres motivated to act as a single coherent unit for a common goal
- Consensus about utility structure to scrutinise efficacy of candidate policies
- Consensus about an overarching description of dynamics driving the process.
- Consensus about who is expert about what, to identify appropriate expert panels
- Expert judgements from disparate panels of experts
- Each component panel informed by complex models & huge data sets

Integrating Decision Support systems

A single, comprehensive probabilistic model is inappropriate

- infeasibly large
- no shared structural assumptions so no centre can 'own' the full joint distribution
- dynamic revisions lead to fast obsolescence

Full technical details in

Coherent Frameworks for Statistical Inference serving Integrating Decision Support Systems *Jim Q. Smith, Martine J. Barons & Manuele Leonelli*, submitted & on arXiv: 1507.07394

Theorem

Theorem:

Suppose an IDSS for a CK class $(\mathbb{U}, \mathbb{D}, \mathbb{S})$ is adequate where $\mathbb{U}$ and $\mathbb{D}$ are arbitrary and $\mathbb{S}$ includes the consensus that the IDSS is delegable, separately informed, cutting and commonly separated at time t. Then it will also be sound and distributed at time t. Furthermore it is common knowledge that the SB's beliefs about each panel's parameter vector $\boldsymbol{\theta}_i$ are the same as those of the corresponding expert panel G_i, $i \in [m]$, for all $d \in \mathbb{D}$ and at any time $t \geq 0$.

Examples of sound and distributive frameworks

- Staged trees
- Bayesian Networks
- Chain event graphs
- Decomposable graphs
- Multiregression dynamic models

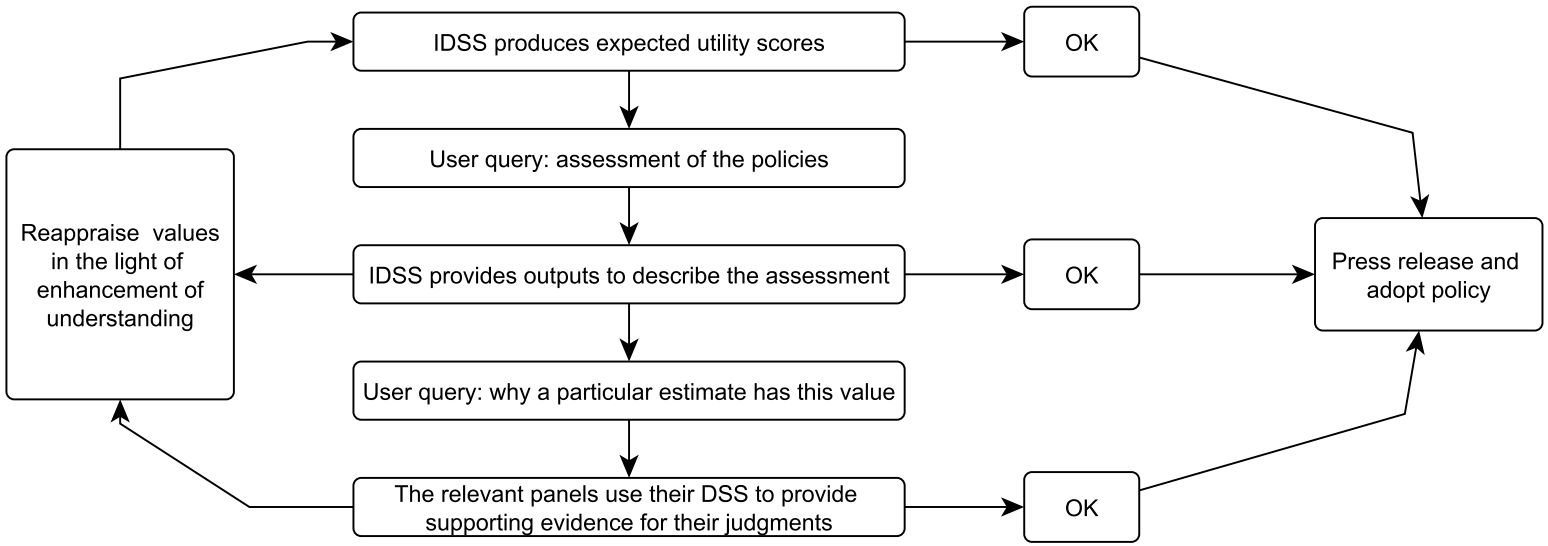


Figure: Manuele Leonelli & James Q. Smith(2015) Bayesian Decision Support for complex systems with many distributed experts Ann Op Res

The National Archives Project

- Soft Elicitation
- Model building
- Model quantification - Data
- Model quantification - SEJ
- Evaluation / feedback
- Software engineering
- Launch & adoption

Soft Elicitation

An essential starting point with any problem is to interact with problem-owners, their advisers, experts and close stakeholders to understand their perspectives, views, values, uncertainties, worldview

Structure the problem

- What are the processes, inputs, outputs, actors, perceptions of cause and effect?
- How do they interact?
- What are the specific objectives, uncertainties, challenges to address?
- How might these be modelled?
- What relevant data and expertise are available?
- How should outputs be presented?
- Iterate

Soft Elicitation with TNA

Problem identification

- Digital preservation is a mission to send messages to the future
- Those messages to be faithfully transmitted, to retain their meaning and to be useful for generations to come.
- Our world is not static. Threats are constantly changing.
- Archives' resources and ability to deal with threats change too and its not always easy to see what to do next.
- Archives all have a long list of good things to do, but the reality is that few of them will ever have the luxury of doing all of them.
- Even the best resourced archives must make choices about how and when to invest in digital preservation.

Soft Elicitation with TNA

A new framework required

A new framework for managing digital preservation risk is required that:

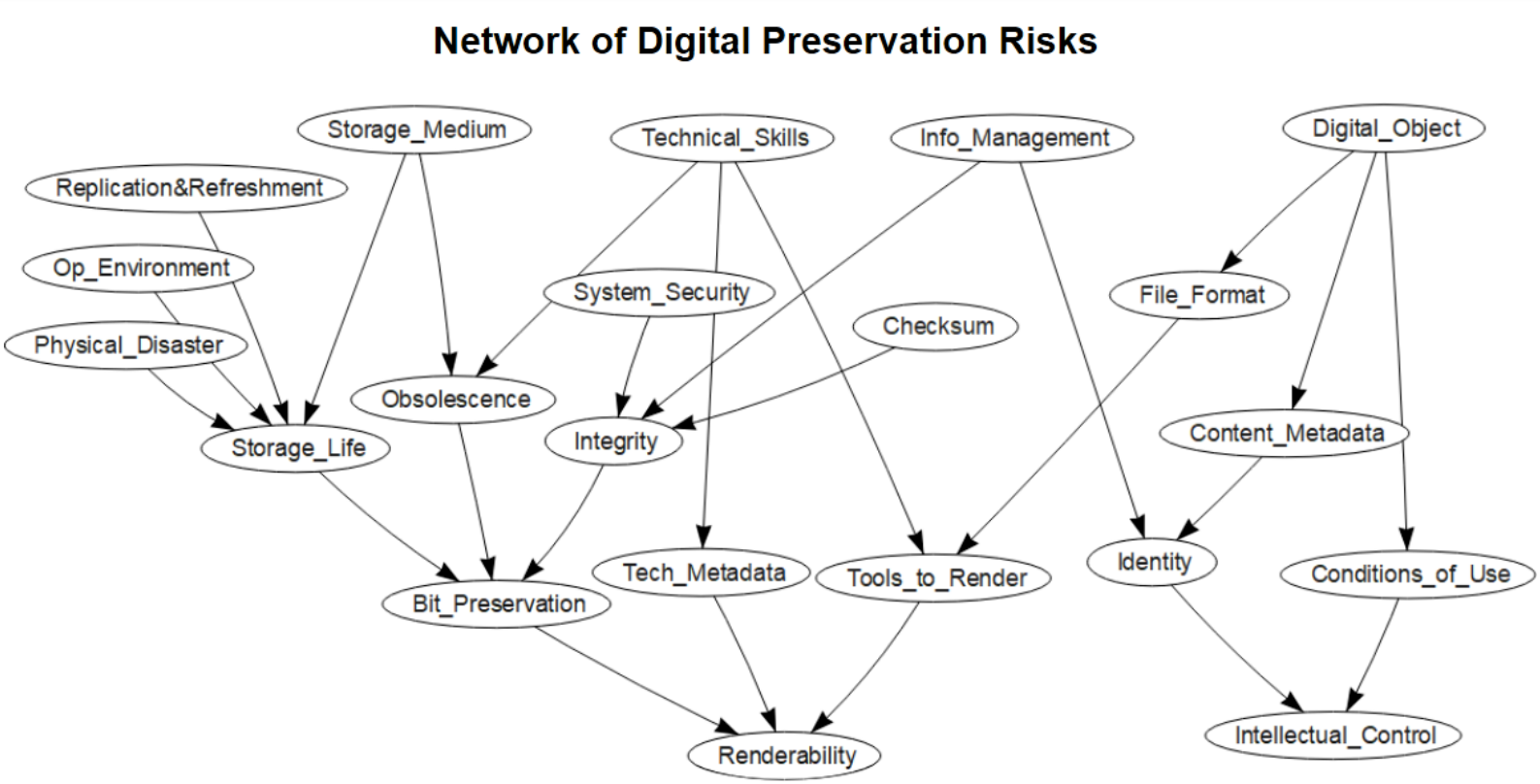
- Describes and explains a complex and interdependent map of risk events, risk management actions and their impact on preservation outcomes.
- Allows archivists to compare and prioritise very different types of threats to the digital archive with potential impact in different areas.
- Operates even where we have limited data or imperfect evidence.

Soft Elicitation with TNA

Workshops

- TNA June 2017 review existing (qualitative) approaches to risk management within digital preservation
- TNA 2017 risk mapping exercise to identify risks within the Digital Repository Infrastructure, and establish relationships between those risks
- November 2018 internal TNA workshop with AS&RU on Bayesian Networks and Structured Expert Judgement
- November 2019 Workshop with TNA and partner archives to identify risks involved in digital preservation and the relationships between them
- Draft network AS&RU with TNA - Dr Thais Fonseca and Hannah Merwood
- January 2020 Formal kick-off and structure review with TNA and partner archives

Network



Utility: Renderability and Intellectual Control

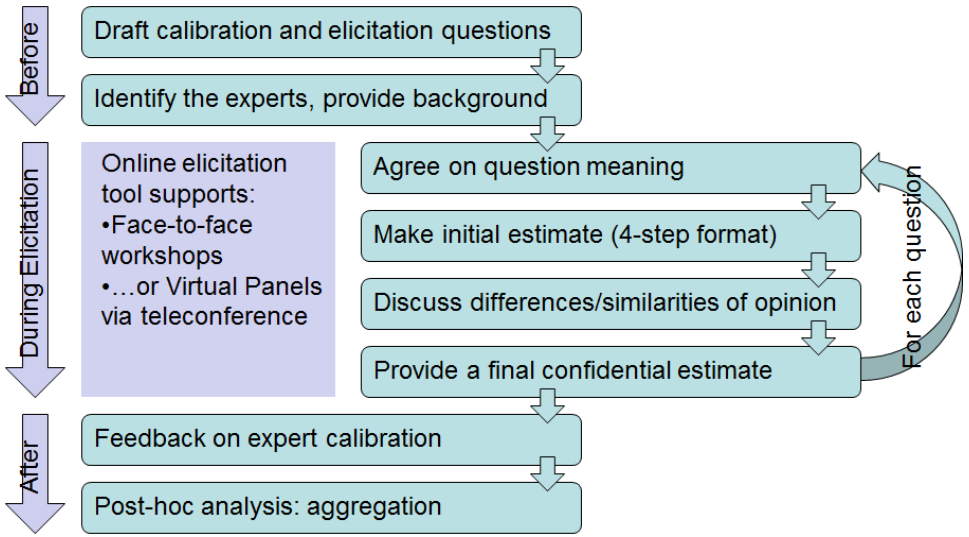
Quantification

Finding relevant data

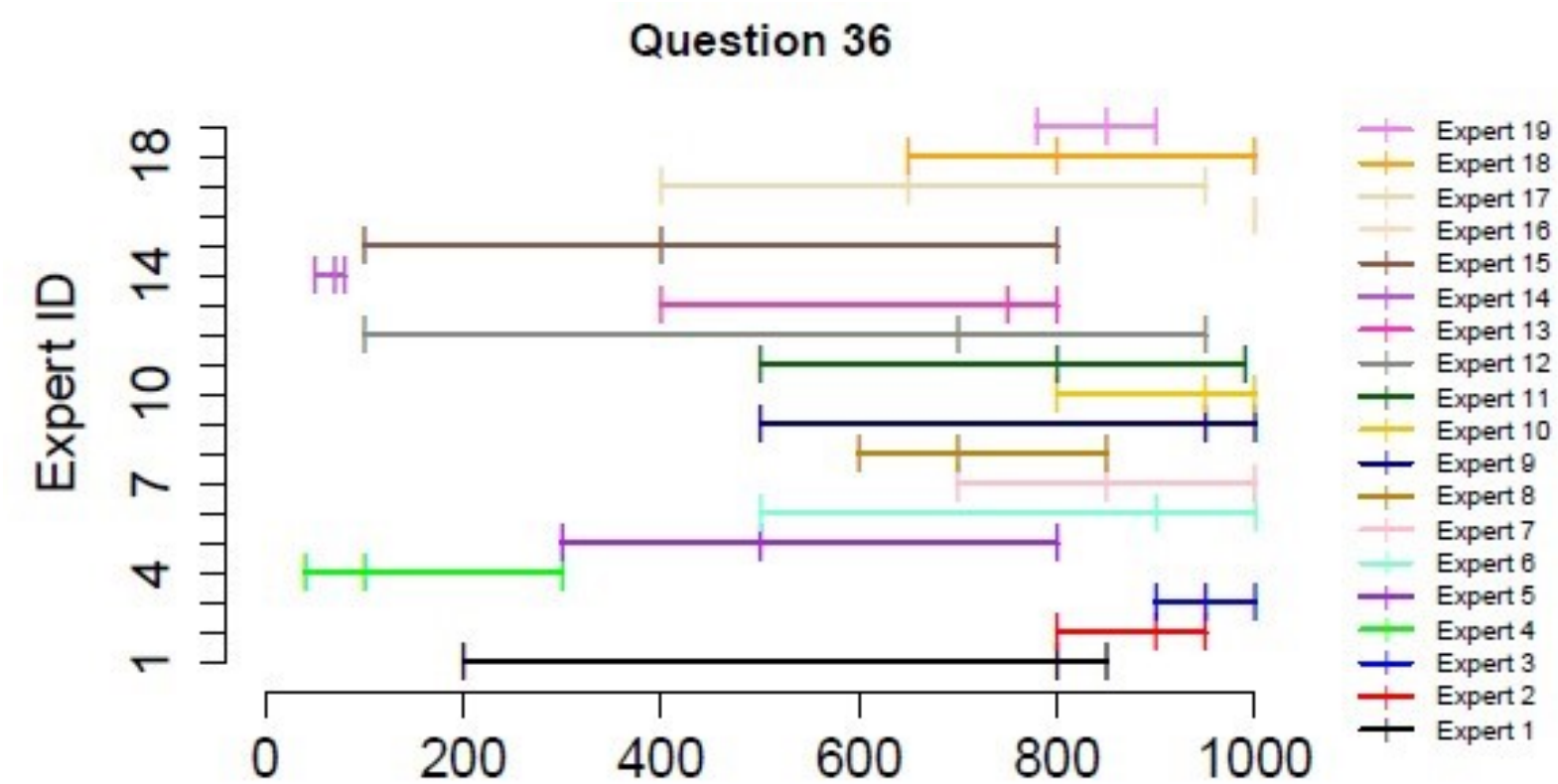
- Node definitions (measurable)
- Surveys
- Previous research
- Available statistics at TNA
- Data granularity - where data does exist they are often related to storage medium and too specific
- Identify data gaps
- SEJ

Structured Expert Judgement - The IDEA protocol

- Investigate, Discuss, Estimate, Aggregate
- Private estimate
- Facilitated Discussion
- Second private estimate
- Aggregation
- Calibration questions
- Questions of interest



SEJ-Range graphs for discussion



Question 36: Out of 1,000 born-digital files, for how many would you expect an archive to know their conditions of use?

A.M. Hanea, M.F. McBride, M.A. Burgman & B.C. Wintle (2018) Classical meets modern in the IDEA protocol for structured expert judgement, Journal of Risk Research, 21:4, 417-433, DOI: 10.1080/13669877.2016.1215346

Safeguarding the nation's digital memory

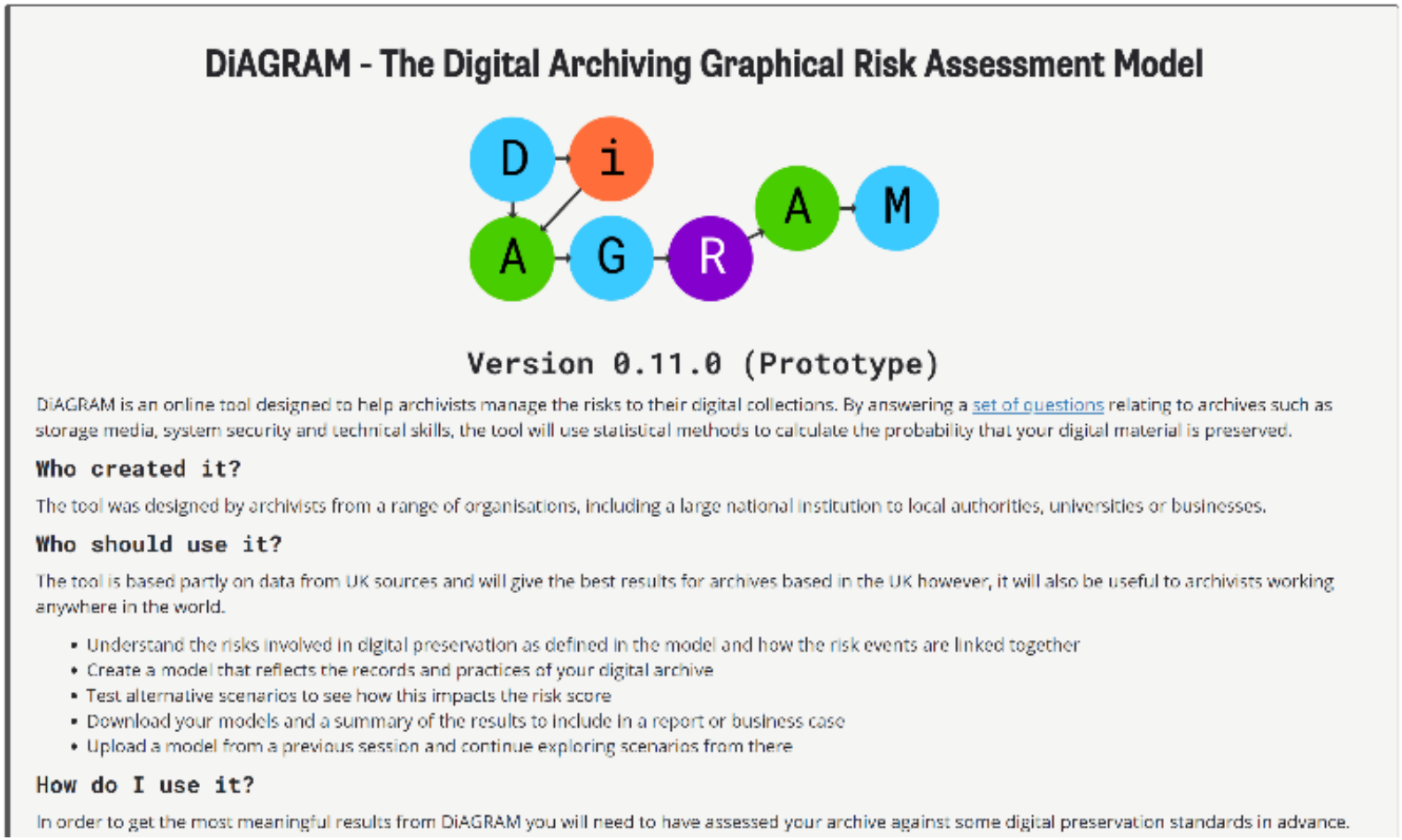


DiAGRAM - The Digital Archiving Graphical Risk Assessment Model

Version 0.11.0 (Prototype)

DiAGRAM is an online tool designed to help archivists manage the risks to their digital collections. By answering a set of questions relating to archives such as storage media, system security and technical skills, the tool will use statistical methods to calculate the probability that your digital material is preserved.

Who created it?

The tool was designed by archivists from a range of organisations, including a large national institution to local authorities, universities or businesses.

Who should use it?

The tool is based partly on data from UK sources and will give the best results for archives based in the UK however, it will also be useful to archivists working anywhere in the world.

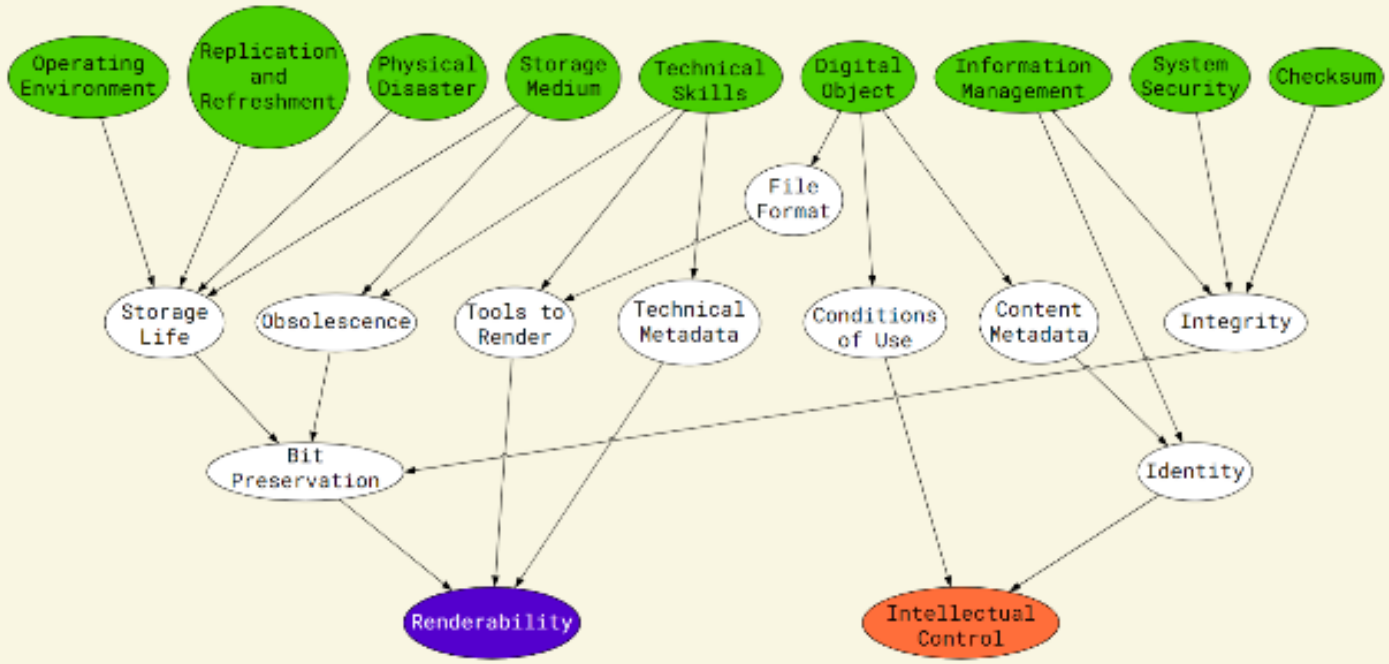
- Understand the risks involved in digital preservation as defined in the model and how the risk events are linked together
- Create a model that reflects the records and practices of your digital archive
- Test alternative scenarios to see how this impacts the risk score
- Download your models and a summary of the results to include in a report or business case
- Upload a model from a previous session and continue exploring scenarios from there

How do I use it?

In order to get the most meaningful results from DiAGRAM you will need to have assessed your archive against some digital preservation standards in advance.

DiAGRAM was built by The National Archives and the University of Warwick with support from the National Lottery Heritage Fund and the Engineering and Physical Sciences Research Council.

DiAGRAM structure



THE NATIONAL ARCHIVES · WARWICK · HERITAGE FUND · UKRI Engineering and Physical Sciences Research Council

Developing DiAGRAM

The National Archive

Workshops to identify variables of interest, granularity, etc.

- The National Archives
- Dorset History Centre
- Gloucestershire Archives
- Transport for London Archives
- University of Brighton Design Archives
- University of Leeds Special Collections

The Utility

- Renderability
- Intellectual control

Digital Preservation Awards 2020 feedback

The judges said ...

- Provides a good summary of the existing DP tools and landscape, and clearly articulates the unique element this project brings (quantitative dimension to risk management).
- Looked at existing DP tools, recognized that a better tool was needed, and found it (Bayesian Networks).
- Outcomes are credible.
- Proof that digital projects are worthwhile in their own right is a benefit perhaps unique to this project.
- COVID created more opportunities to funnel project resources into development. The uniqueness in this project's funding is a stand-out positive impact for potential advocacy for the field

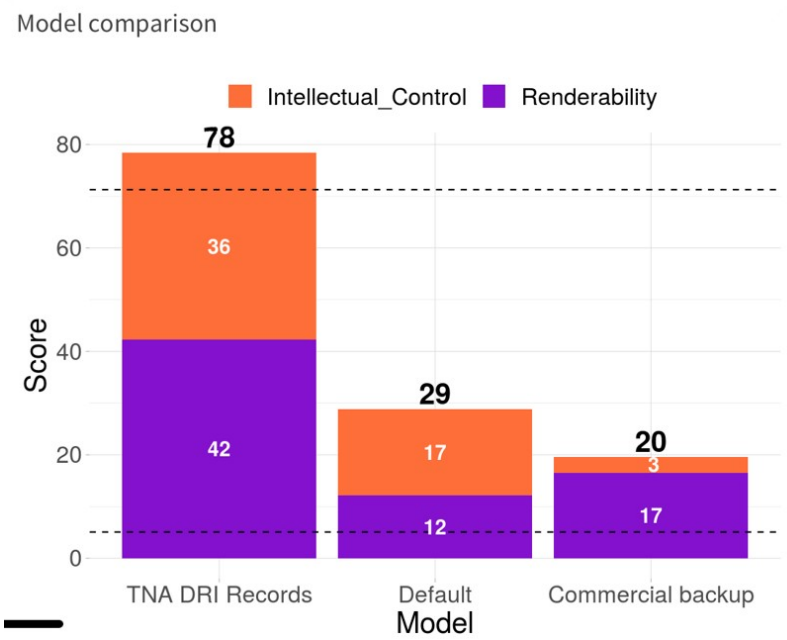
Digital Preservation Awards 2020 feedback

The judges also said ...

- Answers a pressing need with a multidisciplinary approach.
- Clear understanding of audience and what they need from this tool
- The prototype shows clear innovation in the way it displays information and attempts to create quantitative comparisons
- Clear immediate benefits and the starting point for potential new ways of assessing risk
- Very strong, convincing and as far as I can tell entirely new basis for executives to understand the provision of funding, policy and priority to investment in digital preservation, as well as the practical outcomes of preservation actions. This is a step change from anything that has come before.

TNA spending review

- Baseline current digital preservation risk score at The National Archives
- Show current value of digital preservation practice in comparison to model for simple backup
- Model changes to risks scores over various time periods with and without requested funding
- Achieved 33.5% increase in overall funding



Present risk score at The National Archives versus a default average model and just backing up

Acknowledgements

Martine.Barons@warwick.ac.uk
go.warwick.ac.uk/MJBarons

Supported by EPSRC grant EP/K039628/1