Case study: using the DiAGRAM tool for digital preservation

David H Underdown https://orcid.org/0000-0002-8123-4655

Abstract

DiAGRAM (the Digital Archiving Graphical Risk Assessment Model)[1] was initially developed as part of the National Lottery Heritage Fund (NLHF) supported project *Safeguarding the nation's digital memory*. The project partners (The National Archives in the United Kingdom, the Applied Statistics & Risk Unit (AS&RU) at the University of Warwick, Dorset History Centre, Gloucestershire Archives, Transport for London Archives, University of Brighton Design Archives, and University of Leeds Special Collections) collaborated to produce a Bayesian Network (BN) encapsulating digital preservation risks and the interactions between them, and the integrated decision support system (IDSS) built around the BN was given the name DiAGRAM. The work of a postdoctoral researcher at AS&RU was further supported by an Engineering and Physical Sciences Research Council (EPSRC) impact acceleration award. A more detailed examination of the overall project and the creation of the BN and IDSS can be found in the article "Safeguarding the nation's digital memory: towards a Bayesian model of digital preservation risk".[2]

However, that article gives only a basic description of how to use the IDSS and how it:

> "will allow archives to investigate potential mitigations to digital preservation risks based on their own current circumstances, and communicate the relative effectiveness of different strategies (and the costs of different strategies) to relevant decision makers, funders and other stakeholders in an easy-to-understand way. This will allow archives to evidence their requests for support based on a rigorous model which will have been developed using the experience of a wide range of institutions."[3]

This case study examines how DiAGRAM can be used to achieve that objective.

Keywords

Risk modelling, digital preservation, Bayesian networks, advocacy

Project background

Staff at The National Archives (UK) began initial investigations of the potential for the use of Bayesian networks to produce a quantitative model of digital preservation risk in late 2018. This work showed that the approach seemed to have merit, but it also became clear that one major issue was going to be finding appropriate data sources to create the necessary conditional probability tables that allow the quantification of risk. The initial work had also shown that where conventional data sources did not exist we would in theory be able to use the structured elicitation of expert judgement to provide the necessary data, but we lacked the skills and experience to be able to carry this out with appropriate rigour. This led to discussions with the Applied Statistics & Risk Unit at the University of Warwick and development of a bid to the National Lottery Heritage Fund (NLHF) to support the project and allow us to create a network of partner archives to broaden the input into the model in order to ensure that it would be as widely applicable as possible. The archives: Dorset History Centre, Gloucestershire Archives, Transport for London Archives, University of Brighton Design Archives, and University of Leeds Special Collections; represented a cross-section of most types of archive in the UK with experience in carrying out digital preservation (local records offices, university archives and special collections and a corporate archive). We were subsequently also able to include expert contributions from staff at the Digital Preservation Coalition, the Cambridge University Library, BFI (British Film Institute) National Archive and an independent archival consultant.[4]

The project lead at The National Archives was Alex Green and a research assistant, Hannah Merwood, joined on secondment from the Department of Digital, Culture, Media and Sport (DCMS). At AS&RU the work was led by their director, Dr Martine J Barons, and research assistant Dr Thais Fonseca (who was supported by an Engineering and Physical Sciences Research Council Impact Acceleration Award). The prototype integrated decision support system (IDSS) was built by Stephen Krol and Sidhant Bhatia of Monash University who were on a research exchange scheme with the University of Warwick. Subsequent refinement and development of the IDSS was carried out by Jumping Rivers. I wish to acknowledge all their work on the project, and in particular that of Hannah Merwood who carried out much of the initial modelling of our own risk profile at The National Archives without which this article would not have been possible.

The AS&RU team naturally leaned towards the statistical programming language, R,[5] to implement the Bayesian network, and with relatively short project

timelines this led to the choice of Shiny Dashboard[6] for development of the prototype IDSS.[7] Although initially allowing rapid development it proved to come at some cost in difficulties in making the IDSS meet web accessibility requirements. Ultimately Jumping Rivers have redeveloped the IDSS, separating a rewritten front end (in simple HTML with some JavaScript) from an Application Programming Interface (API) connecting to a somewhat refactored back end server with the implementation of the Bayesian network. The API and back end remain in R. The web accessibility and design issues are described in more detail in a long paper presented at the iPRES 2022 conference.[8]

Bayesian Networks and conditional probability

The network of key risks[9] (see Figure 1 below) was developed and agreed by the experts from the various partner archives, facilitated by AS&RU, the definitions used for each node, the states that each can take, and the data source for each, are described in the glossary[10] included in DiAGRAM (the Digital Archiving Graphical Risk Assessment Model) online.

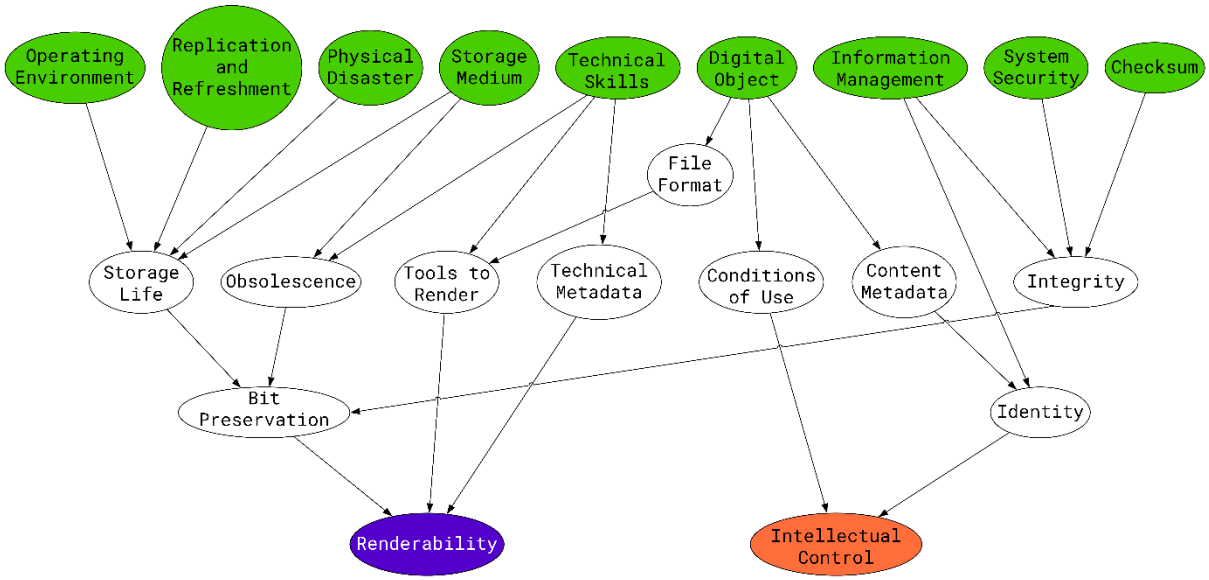

Figure 1 - the network of risks which drives DiAGRAM

In Figure 1, the top row of nodes (green) are the input nodes which must be set up to model the conditions and policies applicable to the individual archive under consideration.

The middle nodes (white) each have an associated conditional probability table (see Table 1 under Advanced Modelling for an example). Depending on the values of the input nodes these conditional probabilities combine and produce the archive's risk scores for Renderability (purple) and Intellectual Control (orange).

Modelling

DiAGRAM has two modelling modes:

1. basic mode using the "Create a model"[11] and then the "Create a scenario"[12] processes. In this mode, when dealing with concepts such as Technical Skills where there is not an obvious objective representation as a percentage then users are invited to make use of their archive's rating for a relevant category of other well-known digital preservation assessment models and these ratings are converted to percentages using a weighting scheme described in more detail below. Creating a scenario allows you to revisit (selected) answers and adjust them to represent a possible future state of your archive so that you can see how risk levels for Renderability and Intellectual Control would change for that scenario. In this mode only input nodes can be set, all middle nodes use the default values of the model.
2. "Advanced customisation"[13] allows direct customisation of both the input nodes and also the conditional probabilities associated with the middle nodes of the BN, rather than relying on only the default values. This allows more fine-grained control of the modelling.

It is possible to mix the modes, creating a basic model through the setup questions and then refining particular nodes using the advanced customisation.

The initial data gathering to create the underlying Bayesian Network determined the median probability values used in the online model, along with 95th percentile (maximum plausible) and 5th percentile (minimum plausible) values. Where you have good evidence to suggest that (for instance) your storage infrastructure is more reliable than the median value it may be reasonable to substitute the 95th percentile value, or of course if you have specific data for your institution relating to a node in the model, to substitute that data for the default median value.

Creating a basic model

The most straightforward way to begin exploring modelling is to use the basic mode, "Create a model". The answers to the various input questions are converted to percentage values (for those which are not straightforward percentages to begin with) and applied to the underlying BN. Since some questions may involve undertaking an assessment against other well-known digital preservation models, or liaising with external departments such as corporate IT a reference version of the questions is provided as a downloadable PDF.[14] Once an initial model has been created, you then have the option to

create scenarios in order to investigate the effect of changes to the inputs. On creating a scenario you are prompted to select the areas of the model you wish to update and will then only be re-presented with the questions relevant to those areas.

The original prototype[7] required direct percentage inputs for all areas which were not straight yes or no answers. However, this posed considerable difficulties for users in determining what an appropriate percentage technical skill level (for example) for their archive actually was and made the selection of input values rather more subjective. Following feedback from initial presentations of the model the present input scheme was developed and user testing undertaken to confirm that this approach was found helpful by users.[8]

The first step in creating a model is to name it, and then click “Next”. You will then enter the input question flow. As you move through the questions, you can also enter notes and comments on your model as you go, which allows you to document assumptions and decisions you incorporate into your model. Your answers to the input questions will be documented by DiAGRAM itself and will be available to you after creating the model so you do not need to record those, but details of your reasoning may be useful.

Questions taking a simple percentage input or similar

Digital Object

The first input question relates to the proportions of each type of digital object in your archive (see Figure 2). The basic screen layout is similar for all questions. Three types are supported, Born Digital, Surrogate (for digitised versions where the original remains the record and has been retained by the archive), and Digitised (where the digital version has become the record, with the original not being retained).

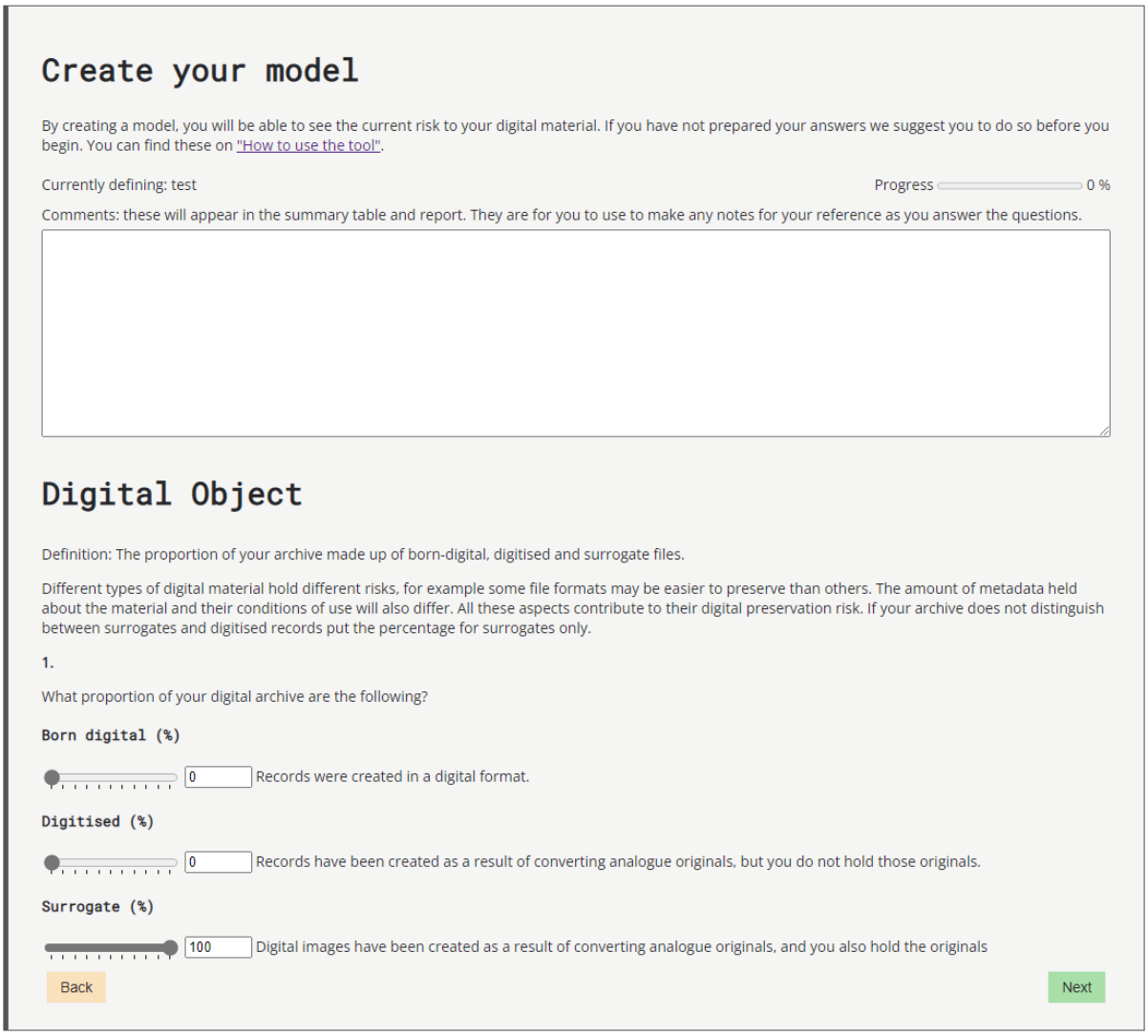

Figure 2 - Data entry screen for the Digital Object node under the basic modelling mode

As this is naturally expressed in percentage terms (it should be based on the number of files of each type, rather than volume in GB) there are simply three selectors presented. You can either click at the appropriate point on the bar, or click into one of the text boxes and enter a percentage directly. The system will check that the percentages always sum to 100% and will adjust other values accordingly as you enter details to ensure that this remains the case. For this question, and all the following ones, there is also a free text entry box for you to record comments which can be used to record details of why you opted for the answer you have chosen. This can be saved with your model. At this stage something to consider is whether you apply the same policies to all types of digital objects. If not, it may be more appropriate to create multiple models. For example, you may tolerate higher risk levels for surrogate material since lost surrogates can be recreated by digitising the originals again, and this may mean you do not maintain off-site copies of this material to reduce costs and

environmental impact. It would not be possible to capture this in a single model, so you should create a model with 100% set for the digital surrogates and set your answers to the remainder of the questions to reflect the policies you have in place for surrogates.

Another consideration here is the types of files you have in your archive, particularly the range of file formats. Based on the data gathered during the original workshop process, if you have born digital material, this feeds through to the File Format node to increase the proportion of files assumed to be of not widely used proprietary file formats (essentially we assume that born digital objects are quite heterogeneous). It also affects our knowledge of Conditions of Use and Content Metadata, for born digital (again based on the original data gathering) we assume that for born digital material we have worse information on those than for Surrogates or Digitised records. So if you actually only hold a narrow range of file formats (or all are open source formats or so common as to be ubiquitous) and have good information on Conditions of Use and Content Metadata you would get a more representative result by describing your material as Digitised rather than as Born Digital.[4]

Storage Media

The next question is again a simple percentage, and looks at your usage of different types of storage media. In the model these are classified as: Type A, Less Stable “Expected lifespan below 10 years or unknown, highly susceptible to physical damage, requires specific environmental conditions and very sensitive to changes, does not support error-detection methods, supporting technology is novel, proprietary and limited. Examples include USB flash drives (memory sticks), floppy disks, SD drives and CD-R discs”; Type B, More Stable “A proven lifespan of at least 10 years, low susceptibility to physical damage, tolerant of a wide range of environmental conditions without data loss, supports robust error-detection methods, supporting technology is well established and widely available. Examples include LTO tapes, Blu-ray discs, enterprise/corporate managed hard drives and CD-ROM discs”; Type C, Outsourced Data Storage “An external company is responsible for our digital storage. Examples include Amazon Simple Storage Service, Microsoft Azure Archive Storage and Google Cloud Storage”. For this question a consideration is the likelihood that there could be material on a variety of media (particularly media which would be classified as Type A) mixed in with paper materials that you hold and as yet unprocessed. Selecting Type C also has an impact on the question following, rendering your answer on Replication and Refreshment largely irrelevant to the model as it is assumed that the cloud provider will be carrying out ongoing

refreshment work and effectively maintains multiple copies. Similarly the Physical Disaster risk is also rendered largely moot.

Replication and Refreshment

The next questions (Replication and Refreshment) refer to your archive's policy on replicating stored material (that is having at least one additional copy) and ensuring that these copies are independent, (that is the copies are on different media and fresh copies made as media ages). Again these are simple percentages. You should ensure that you fully consider all media included in your answer to the previous question, taking account of media from which original records have not yet been copied.

Operating Environment

Next we consider the archive's Operating Environment. The two questions in this section relate to the archive's policies on protection in the event of a physical disaster affecting the archive's primary storage location. Currently these specifically relate to protections against flood risk. The first question asks about the percentage of digital materials for which an offsite copy is kept, while the second asks if you have adequate mitigations in place to at the primary storage location to protect against flood damage (this is a straight yes/no question, or not applicable where copies are all offsite).

Flood Risk

Then the model examines the actual risk of flood at the archive's location. Due to the tool's development in the UK, users will be directed to use the UK Government's postcode based flood risk checking tool https://www.gov.uk/check-long-term-flood-risk. This classifies locations as either Very Low, Low, Medium, or High for either surface water flooding (flash flooding) or sea/river flooding. You should choose the higher of the two. The flood risk tool gives a range of probabilities for each band, but within DiAGRAM we take the median of each band as the risk. The bands are 0-0.1% (treated as 0.05% in DiAGRAM), 0.1%-1% (0.5%), 1%-3.3% (2%) and 3.3%-100% (5%) annually.

Archives outside the UK will need to seek local data sources. If different banding is used you may need to adjust your model using the "Advanced modelling" process in order to accurately reflect the risk. Alternatively, for a quick view of the potential impact of adjusting this figure you could create a model and related scenario using values in DiAGRAM which fall either side of the stated value.

We will be happy to discuss this, along with the potential for modelling other forms of physical risk within DiAGRAM.

Checksum

The next question relates to the percentage of files for which your archive has checksum information, and whether this was obtained before or after material was transferred to the archive. The ideal is to have a checksum which was generated on the donor's system as you can then absolutely assert that you have the identical file. Failing that you should generate a checksum as soon after receipt as possible, but this still leaves the possibility that the file was accidentally changed during the transfer. Finally there is the case where you have no checksum at all or it was generated a long time after receipt in which case there is a much higher risk that the file has become corrupted (or otherwise altered) before the checksum was calculated. Once again you should take account of material that has not yet been extracted from its original media. This question again simply takes percentage values and is constrained to ensure that the total entered always makes 100%.

If you answer "No" for the majority of your material it will make the answers to the following questions on System Security somewhat irrelevant as you can never guarantee the integrity of the material, regardless of how good your security is.

Questions using weighting scheme to convert to percentages

The remaining questions using the weighting approach mentioned at the start of this section to convert the answers to percentage values.

System Security

Firstly you are asked about the security of your archive systems. This question has four sub-sections:

1. Accreditation, a multiple choice question which will see the input percentage reach a maximum of 70% of the total for the section
    - None (0%)
    - Cyber Essentials (10%)[15]
    - Cyber Essentials Plus (40%)[15]
    - ISO 27001 (70%)
2. Penetration testing, again a multiple choice question which carries up to 15% of the total for the section
    - No test (0%)
    - Tested, but critical issues remain to be resolved (5%)
    - Tested, but severe issues remain to be resolved (10%)
    - Tested, no, or only minor issues, outstanding (15%)

3. Assessment against the Control functional area of the NDSA Levels of Preservation[16] which carries up to 10% of the total for the section
    - Not achieved (0%)
    - Level 1 (2%)
    - Level 2 (4%)
    - Level 3 (7%)
    - Level 4 (10%)
4. Virus checks, has all material been virus checked and a record made of the outcome. This may be affected by unprocessed material. This carries up to 5% of the total for the section.
    - No (0%)
    - Yes (5%)

Cyber Essentials and Cyber Essentials Plus are IT security schemes backed by the UK government, primarily aimed at smaller and medium enterprises.

Information Management

Information Management comprises three subsections, though slightly differently organised and with the third examining two related areas but with similar weighting:

1. Assessment against the Metadata functional area of the NDSA Levels of Preservation which carries up to 28% of the total for the section
    - Not achieved (0%)
    - Level 1 (7%)
    - Level 2 (14%)
    - Level 3 (21%)
    - Level 4 (28%)
2. Assessment against the Content functional area of the NDSA Levels of Preservation which carries up to 16% of the total for the section (in effect it is only achieving at least Level 2 for this functional area that matters)
    - Not achieved (0%)
    - Level 1 (8%)
    - Level 2 (16%)
    - Level 3 (16%)
    - Level 4 (16%)
3. Assessment against Section I – Content Preservation and Section J – Metadata Management of the DPC RAM.[17] Each carries 28% of the total for a maximum of 56% of the total across the paired sections, each is scored as follows
    - Minimal awareness (0%)

- Awareness (7%)
- Basic (14%)
- Managed (21%)
- Optimized (28%)

Technical Skills

The final section is an assessment against ten skills selected from the DigCurV[18] skills matrix. Each skill is worth up to 10% of the section total according to the following scale:

1. None (0%)
2. Basic – "is aware of" (3%)
3. Intermediate – "understands" (6%)
4. Advanced – "is able to" (10%)

The skills selected (which seemed the most relevant to the rest of the risk model) are:

KIA 1.9	Apply appropriate technological solutions
KIA 1.12	Digital preservation standards
KIA 1.15	Information technology definitions and skills
KIA 1.16	Select and apply digital curation and preservation techniques
KIA 3.4	Continuously monitor and evaluate digital curation technologies
KIA 5.1	Data structures and types
KIA 5.2	File types, applications and systems
KIA 5.3	Database types and structures
KIA 5.4	Execute analysis of and forensic procedures in digital curation
PQ 3.9	Translate current digital curation knowledge into new services and tools

Baseline model

By following through these questions you will build a baseline model for the current state of your archive. To get an idea of the relative performance of your archive in terms of managing digital preservation risk you can compare this baseline to the two built-in models provided within DiAGRAM. To do this use the View results page.

These model simple commercial backup as the low end, helping us to answer the question of what digital preservation offers over and above backup. This model scores 38% for Renderability (so in the short term at least it is likely you will be able to open your files), but just 6% for Intellectual Control: backups don't

usually carry information about copyright, and if you only want to restore a single file it may not be straightforward to locate it. Nor will you have technical information about the files and the file formats included in the archive.

At the high end we model what's described as "Established National Archive", the position that may be achieved by a large, relatively well-resourced, archive that has been undertaking digital preservation actions for some years. In this model the score is 61% for Renderability and 68% for Intellectual Control. There is little scope for improving this score within the simple modelling paradigm, they could improve their security score very slightly by doing more to monitor access logs, and there are still improvements they could make (inevitably) to their metadata. More substantial improvements would require the use of Advanced Modelling and being able to demonstrate that (for instance) the risk of obsolescence they are facing is lower than that at the median level determined by structured elicitation of expert judgement that was used to derive the values included in the standard model.

In both cases you can view the assumptions used in these models by going to the View results page and clicking "Show answers".

Create a scenario

To create a scenario[12] you essentially re-answer the original input questions, revising your answers to reflect your desired state of the digital archive, rather than its current, baseline, state.

To begin, select some or all of the original input areas on the first screen, and click "Create a scenario". Then give it a name (it will probably be useful to include a high level indication of what the scenario is designed to show, for example "Improve technical skills"), click "Next" and you will be redirected into the input questions flow, but showing only the selected questions relevant to your scenario. Again, you can also add comments in order to document your assumptions or anything else in relation to the scenario that you might subsequently find useful.

Having created your scenario, the "View results" page will show it grouped with the original model to make it easy to compare how the scores for Renderability and Intellectual Control have changed.

You can then create further scenarios, either returning to your original baseline model each time in order to compare different potential courses of action to see which gives you the greatest improvement, or layer further changes on top of a

previous scenario in order to see how you could use incremental changes in different areas to drive improvements in your scores.

Scenario modelling considerations

To get the most out of this scenario modelling you should also ensure that you understand the budgetary implications of different scenarios: if one scenario gives only a marginally smaller improvement than another but is considerably cheaper to implement you may find it easier to make the case for the cheaper course of action.

In addition to modelling your desired state of affairs it will often be useful to model a counterfactual: what would happen if you are unable to carry out the proposed actions due to lack of resources. In many cases this would not simply mean that risk levels remained the same, in fact risks are likely to increase due to factors such as increased obsolescence and technical skills not being kept up-to-date.

In undertaking our own modelling at The National Archives we approached this by assuming that a reduction would be proportional to the real terms reduction in funding resulting from a flat cash settlement (that is, we would continue receiving our existing level of funding from central government with no adjustment for inflation). In our case (at the time of modelling) this was expected to be equivalent to a year-on-year 5% annual reduction. This will compound over time, so if we take our initial funding level as 100%, then after one year our funding will be 100 * 0.95 = 95%, but after two years at flat cash the current value of the second year's funding would be 95% of the first year's value, that is 100 * 0.95 * 0.95 = 90.25% of our original funding. This continues depending how far into the future you are modelling, with each year being (in this example) 95% of the value of the previous year. In accountancy terms this is called a discount rate. A number of factors affect the most appropriate rate to use, so consult a management accountant to determine discount rate for your modelling, based on considerations such as the rate of inflation and the accounting standards in use at your organisation.[19,20]

For technical skills you may be able to use the weightings described above to create a set of technical skills which would give 90% of your original value for technical skills (most simply, if your archive started actually at 100% for technical skills, set 9 of the skills to Level 4 and the final one to Level 0, however if your skills started at a lesser value you would first need to calculate 90.25% of your original value and then see if you could craft an appropriate set of skills). However, since obsolescence is not one of the input nodes there is no simple

way for you to amend the value using the basic modelling approach, to amend these nodes we need to introduce advanced modelling using the Advanced customisation flow. [13]

Advanced Modelling

When creating an advanced model you can choose to start either from an existing model (including any models created by the basic modelling flow) or start from scratch. If you opt to start from scratch make sure you set suitable values for all input nodes. If you start from an existing model you can choose to make your advanced model a scenario of the original model or a separate model in its own right. The main difference is that a scenario will be grouped with the original base model under "View results" rather than shown separately. Having made that choice and named the resulting model or scenario, click next and you be presented with the modelling screen.

You can now choose a node for which to update the probability table. You can either click on the node directly on the network view, or use the dropdown which lists all nodes. Having selected a node it, along with its parent and child nodes (and their parents and children if relevant), will be highlighted in the network diagram, with unrelated nodes greyed out. The probability table for the node will be shown below, usually lower left, though the exact screen layout will depend on the screen width.

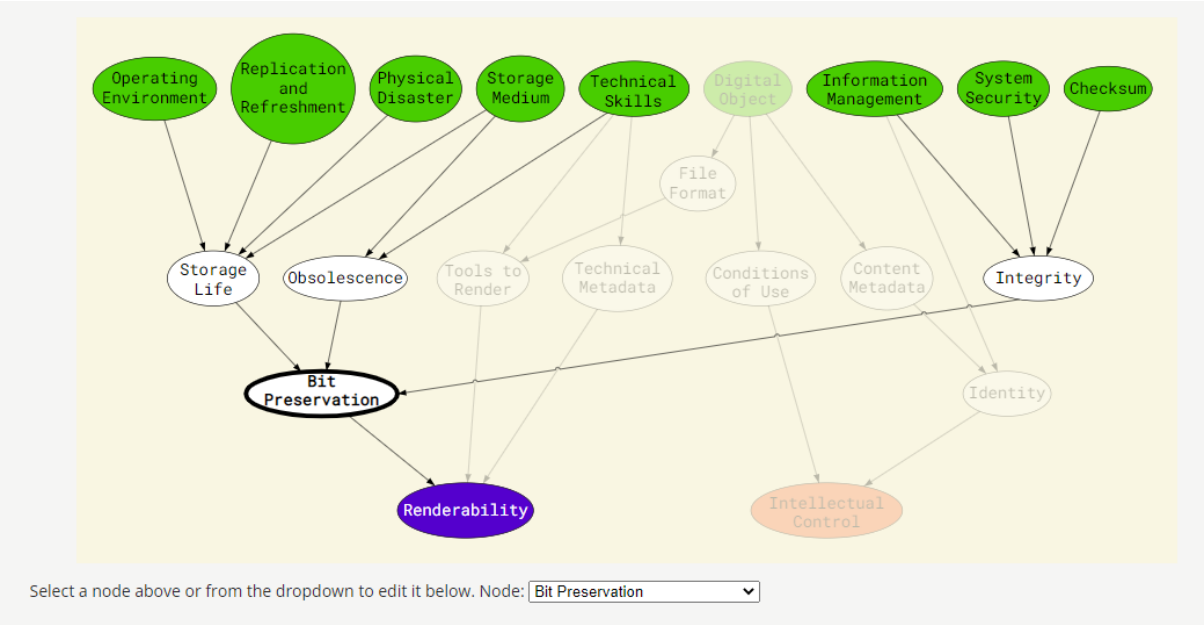

Figure 3 - Bit preservation node selected, unrelated nodes greyed out

You can then edit the probabilities for the node values. There will be a row for each possible combination of ancestor nodes and you are entering probabilities

for each possible state of your chosen node given that combination. The total probabilities for each row of the table must total 1.

Table 1 – conditional probability table for the Bit Preservation node

	Integrity	Obsolescence	Storage Life	Yes	No
1	Yes	Yes	Yes	0.0000	1.0000
2	Yes	Yes	No	0.0000	1.0000
3	Yes	No	Yes	1.0000	0.0000
4	Yes	No	No	0.0000	1.0000
5	No	Yes	Yes	0.0000	1.0000
6	No	Yes	No	0.0000	1.0000
7	No	No	Yes	0.7158	0.2842
8	No	No	No	0.0000	1.0000

Having selected a node there is also a link to the relevant glossary entry (which will open in a new tab)[10]. To fully understand the table you may also need to view the glossary entries for the ancestor nodes to full understand the table. For example, looking at the Bit Preservation node, then for the ancestor nodes Integrity and Storage Life then the desirable node state for those is Yes, but Obsolescence then No would be the preferred state. Taking that into account it becomes clearer why Obsolescence being Yes immediately forces Bit Preservation to No, and the same is also true for Storage Life being No. However, Integrity being No with Obsolescence No and Storage Life Yes gives us probabilities for Bit Preservation of 0.7158 Yes and 0.2842 No.

To determine appropriate alternative values to use when using Advanced customisation you may have a specific source of evidence from your own archive to give alternative probabilities. Another approach is to consider the ranges produced by the elicitation process. The values used in DiAGRAM are the median probabilities from the elicitation. If, for example, you believe your storage system is more reliable and less prone to obsolescence than average you could take the 95th percentile values from the elicitation rather than the median.

Another use could be in fine tuning input values, for example those which use the NDSA Levels[16] or DPC RAM,[17] you may feel that you only just miss making

the next level up (having some but not all of the requirements in place). So if on the Metadata functional area you have met the requirements for Level 3, and have some of the Level 4 requirements in place you might wish to increase your score on Information Management by 4% to reflect that partial compliance. Alternatively that could be your development scenario, again you won't fully meet Level 4 in the time period for which you are modelling, but will have made progress towards it and it would be fair to reflect this in reduced risk to your digital archive.

Support for advocacy and building business cases

Models and scenarios (whether created via the simple of advanced process) can be saved and downloaded from the "Download a report" page, and a PDF slide pack can also be generated. The prototypes allowed files to be exported in R's own .rds format, however for the live version we have switched to JSON (JavaScript Object Notation)[21] which should be easier to support in the long-term with less risk of changes in the application meaning that the .rds files from one version are not readable subsequently. You can also download a CSV file to allow you to do your own further analysis and create your own charts (for instance in Excel). The charts on the "View results" page are static images so can be saved by right-clicking and choosing "save image as…" for inclusion in your own documents.

When saving you can deselect models and scenarios you do not wish to include. It is important to note that nothing is saved on the server, so if you do not download your own models and scenarios they will be lost. This has the advantage that no-one at The National Archives (or on the support team at Jumping Rivers) can see any of your modelling so information remains confidential.

If you save data as JSON you can then use the "Upload previous models" page to reload the data and generate further models and scenarios using your previous work as a basis.

Data may also remain in your browser's local storage for 24 hours and this means that you will be able to review your modelling for this period. No data is stored on the servers so modelling remains confidential.

The PDF slide pack includes all charts that are on the "View results" page and for each selected model or scenario it will include the answers to the input questions for simple models or information on changed conditional probability tables for advanced models. Note that due to the complexities of PDF there are

some remaining issues with accessibility of information in that form, however everything is available elsewhere within DiAGRAM which we believe is accessible, or alternative formats can be provided on request.

The statistical rigour of the model provided by the partnership with AS&RU, and the breadth of expert input due to the range of archives partners in the original project, and additional digital preservation experts who took part in the elicitation process ensure that the model draws on a wide range of experience and is broadly applicable (though the data is most relevant to UK-based archives). The National Archives have used DiAGRAM modelling as evidence in their own submissions in the UK government's spending review process in 2020 and 2021 where it proved extremely useful in conveying the need for further investment in systems to reduce risk to digital public records.

Conclusion

DiAGRAM provides a means for archives to establish the current risk levels to their digital holdings and then investigate potential scenarios for reducing those risks, or demonstrate how risks will increase without appropriate investment.

Simple modelling can be undertaken using a "wizard"-like process of guided questions which make use of a variety of established digital preservation practice such as the NDSA Levels of Preservation and the DPC's Rapid Assessment Model. Advanced modelling, which involves directly editing the probability data underlying the network diagram at the core of the model is also available.

Models and scenarios prepared using either approach are not stored within the online tool in order to maintain confidentiality of assessments, but can be downloaded by their creator and then re-uploaded if further modelling is required. Images of the charts created by the tool and a PDF slide pack of information on selected models and scenarios can also be downloaded to help with advocacy work. A CSV file can also be downloaded for simple models to allow users to create their own additional charting or analysis.

References

1. Applied Statistics & Risk Unit, the University of Warwick/The National Archives/Jumping Rivers. *DiAGRAM Version 1.0.0*. https://diagram.nationalarchives.gov.uk [accessed 26 October 2022]
2. Barons M, Bhatia S, Double J, Fonseca T, Green A, Krol S, et al. (2021) Safeguarding the nation's digital memory: towards a Bayesian model of digital preservation risk, *Archives and Records*. 2021; 42(1): 58-78.

https://doi.org/10.1080/23257962.2021.1873121 [accessed 22 November 2022]
3. The National Archives. *Safeguarding the nation's digital memory.* https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/safeguarding-the-nations-digital-memory/ [accessed 2 September 2022]
4. Merwood H, Green A, Underdown DH, Barons MJ, Fonseca T. The National Archives. *Report of the digital preservation expert elicitation workshop held online 28 & 29 April 2020*. 2022. https://www.nationalarchives.gov.uk/about/our-research-and-academic-collaboration/our-research-projects/open-access-research-from-our-staff/report-of-the-digital-preservation-expert-elicitation-workshop-held-online-28-and-29-april-2020/ [accessed 18 November 2022]
5. The R Foundation, "What is R?", *The R Project for Statistical Computing*. Not dated. https://www.r-project.org/about.html [accessed 22 November 2022]
6. R Studio, inc. Home. *shinydashboard* https://rstudio.github.io/shinydashboard/index.html [accessed 22 November 2022]
7. Applied Statistics & Risk Unit, the University of Warwick/The National Archives. *DiAGRAM Version 0.8.0*. https://nationalarchives.shinyapps.io/tna_gui/ [accessed 2 September 2022]
8. Underdown D, Leigh A, Descheemaeker P. Making Risk Modeling Accessible with DiAGRAM In: *iPres2022, Proceedings of the 18th International Conference on Digital Preservation, 12-16 September 2022, Glasgow, United Kingdom*. Glasgow: Digital Preservation Coalition; 2022. http://doi.org/10.7207/ipres2022-proceedings#page=174 [accessed 18 November 2022]
9. Applied Statistics & Risk Unit, the University of Warwick/The National Archives/Jumping Rivers. Learn about DiAGRAM. *DiAGRAM Version 1.0.0*. https://diagram.nationalarchives.gov.uk/definitions.html [accessed 26 October 2022]
10. Applied Statistics & Risk Unit, the University of Warwick/The National Archives/Jumping Rivers. Glossary. *DiAGRAM Version 1.0.0*. https://diagram.nationalarchives.gov.uk/glossary.html [accessed 26 October 2022]

11. Applied Statistics & Risk Unit, the University of Warwick/The National Archives/Jumping Rivers. Create a model. *DiAGRAM Version 1.0.0*. https://diagram.nationalarchives.gov.uk/model.html [accessed 26 October 2022]
12. Applied Statistics & Risk Unit, the University of Warwick/The National Archives/Jumping Rivers. Create a scenario. *DiAGRAM Version 1.0.0*. https://diagram.nationalarchives.gov.uk/scenario.html [accessed 26 October 2022]
13. Applied Statistics & Risk Unit, the University of Warwick/The National Archives/Jumping Rivers. Advanced customisation. *DiAGRAM Version 1.0.0*. https://diagram.nationalarchives.gov.uk/advanced.html [accessed 26 October 2022]
14. Applied Statistics & Risk Unit, the University of Warwick/The National Archives/Jumping Rivers. Question set PDF. *DiAGRAM Version 1.0.0*. https://diagram.nationalarchives.gov.uk/www/final_questions.pdf [accessed 26 October 2022]
15. National Cyber Security Centre. About Cyber Essentials. Not dated. https://www.ncsc.gov.uk/cyberessentials/overview [accessed 22 November 2022]
16. National Digital Stewardship Association. *Levels of Digital Preservation, version 2.0*. 2019. https://ndsa.org/publications/levels-of-digital-preservation/ [accessed 18 November 2022]
17. Digital Preservation Coalition. *Rapid Assessment Model version 2.* March 2021. https://www.dpconline.org/digipres/implement-digipres/dpc-ram [accessed 21 November 2022]
18. DigCurV. Skills and Competency Levels. 2013. https://digcurv.gla.ac.uk/skills.html [accessed 22 November 2022]
19. Office of Budget Responsibility. *Discount rates*, July 2011. https://obr.uk/box/discount-rates/ [accessed 21 November 2022]
20. KPMG. *Determining appropriate discount rates in an uncertain environment: Exploring the challenges for companies in forming accounting estimates*, November 2020. https://home.kpmg/uk/en/home/insights/2020/11/determining-appropriate-discount-rates.html [accessed 21 November 2022]
21. *Introducing JSON*. Not dated. https://www.json.org/json-en.html [accessed 21 November 2022]

Author and organisation information

David H Underdown is a senior digital archivist at The National Archives where he has been employed since 2005. He was part of the core project team for the *Safeguarding the nation's digital memory* project, supported by the National Lottery Heritage Fund and the Engineering and Physical Sciences Research Council, which developed the DiAGRAM risk modelling tool.

Prior to joining The National Archives he read a for BSc (Hons) degrees in mathematics at Imperial College from 1995 to 1998 and subsequently worked in information technology for an insurance company for several years, all of which provided useful background for the statistical content of the DiAGRAM tool.

The National Archives are a non-ministerial department, and the official archive and publisher for the UK Government, and for England and Wales. They are the guardians of over 1,000 years of iconic national documents.

They are expert advisers in information and records management and are a cultural, academic and heritage institution. We fulfil a leadership role for the archive sector and work to secure the future of physical and digital records.