Evaluating Assessment Frameworks in Digital Curation

Christoph Becker, Emily Maemura, Nathan Moles, Jess Whyte, Hannes Kulovits, Steve Marks
April 2016 / June 2017 / July 2018
Author contact: christoph.becker@utoronto.ca

Summary:
Assessment of organizational digital curation capability and maturity plays an important role as a diagnostic instrument to support planning and prioritize areas of an organization's capabilities to improve. The last decade has seen a surge in the number of models and tools that aim to support digital curation practitioners in this process. The central role of these frameworks in strategic decision making suggests that they are worth studying, but the frameworks themselves have not been the subject of evaluation comparable to that expected of organizations themselves.

This report describes a study of digital curation assessment frameworks in theory and practice. The report presents a case study methodology to support the systematic study of such frameworks in use and describes the application of this method with practitioners at an academic library and a national archive. This presents a rich set of evidence about four cases, collected and organized to facilitate the analysis of two frameworks. A comparative case study analysis characterizes these frameworks across multiple real contexts of use to facilitate their evaluation. The evaluation method is described and documented in detail to facilitate its replication in other contexts and to other frameworks.

This evaluation of frameworks in practice is based on theory and frameworks from design science research methodologies and complemented by an analysis of their design and the process followed in designing them. The findings identify opportunities to improve these frameworks and raise questions for further investigation into the design of such frameworks.

Copyright by the Authors.
This work was supported by the Vienna Science and Technology Fund (WWTF) through the project BenchmarkDP (ICT12-046).
1 INTRODUCTION ........................................................................................................... 3
2 BACKGROUND ............................................................................................................ 4
  2.1 ORGANIZATIONAL ASSESSMENT IN DIGITAL CURATION ..................................... 4
  2.2 THE DESIGN OF ASSESSMENT FRAMEWORKS ....................................................... 5
  2.3 EVALUATING ASSESSMENT FRAMEWORKS .......................................................... 6
3 A METHOD FOR EVALUATING ASSESSMENT FRAMEWORKS ................................... 8
4 THE SELECTION OF FRAMEWORKS FOR EVALUATION ........................................... 10
  4.1 ORGANIZATIONAL CONTEXT .............................................................................. 10
  4.2 SELECTION CRITERIA AND FRAMEWORKS ......................................................... 10
  4.3 DPCMM: THE DIGITAL PRESERVATION CAPABILITY MATURITY MODEL ............. 11
  4.4 AIDA: ASSESSING INSTITUTIONAL DIGITAL ASSETS ........................................ 13
5 AN ASSESSMENT PROCEDURE ............................................................................... 15
6 PROCESS EVALUATION ............................................................................................ 16
  6.1 PROCESS EVALUATION FOR DPCMM ................................................................. 17
  6.2 PROCESS EVALUATION FOR AIDA .................................................................... 18
7 DESIGN EVALUATION ............................................................................................... 20
  7.1 DESIGN EVALUATION FOR THE DPCMM ............................................................ 20
  7.2 DESIGN EVALUATION OF AIDA ........................................................................ 22
8 USAGE EVALUATION ............................................................................................... 25
  8.1 CASE STUDY DESIGN ............................................................................................ 25
  8.2 DATA COLLECTION ............................................................................................... 26
  8.3 ANALYSIS AND REPORTING ............................................................................... 27
  8.4 CASE 1: DPCMM IN USE AT ASA ....................................................................... 27
  8.5 CASE 2: DPCMM IN USE AT UTL ...................................................................... 30
  8.6 CASE 3: AIDA IN USE AT ASA .......................................................................... 34
  8.7 CASE 4: AIDA IN USE AT UTL .......................................................................... 37
9 ANALYSIS OF ASSESSMENT PROJECTS .................................................................. 41
  9.1 RESOURCES AND SKILLS .................................................................................... 41
  9.2 VALUE AND USEFULNESS ................................................................................... 41
  9.3 ACCURACY AND RELIABILITY ............................................................................. 42
  9.4 DIFFICULTIES AND OPPORTUNITIES ................................................................ 42
10 ANALYSIS OF ASSESSMENT FRAMEWORKS ......................................................... 43
  10.1 ANALYSIS OF THE DPCMM IN USE ................................................................. 43
  10.2 ANALYSIS OF AIDA IN USE .............................................................................. 45
11 SYNTHESIS OF EVALUATION .............................................................................. 47
  11.1 EXPLAINING INCIDENTS USING DESIGN EVALUATION RESULTS .................. 47
  11.2 INTERPRETING AND APPLYING CRITERIA AND TERMINOLOGY ...................... 47
  11.3 CONCEPTIONS OF MATURITY ........................................................................... 49
12 CONCLUSIONS ........................................................................................................ 51
  12.1 LIMITATIONS ..................................................................................................... 51
13 SUMMARY AND OUTLOOK .................................................................................... 53
REFERENCES .................................................................................................................. 54
APPENDIX A: PRE-ASSESSMENT SURVEY ................................................................. 58
APPENDIX B: POST-ASSESSMENT INTERVIEW QUESTIONS .................................... 59
1 Introduction

Assessment of organizational digital curation capability and maturity plays an important role as a diagnostic instrument to support planning and prioritize areas of an organization’s capabilities to improve. The last decade has seen a surge in the number of models and tools that aim to support digital curation practitioners in this process. The central role of these frameworks in strategic decision making suggests that they are worth studying, but the frameworks themselves have not been the subject of evaluation comparable to that expected of organizations themselves. The question then arises: How can existing assessment frameworks be studied and evaluated systematically?

To address this challenge, this report presents a comprehensive evaluation method to study the design and use of assessment frameworks in theory and practice. The first part is an analytic evaluation of the framework’s design and a review of its design process. The second component is an empirical methodology for characterizing assessment frameworks through case studies analyzing their use. The method provides a replicable and structured approach to study a given framework to allow insights into how it influences the assessment process.

We applied this evaluation method to two frameworks. The analytic evaluation component provides a structured analysis of the design and design process of each framework. The empirical component provides a rich set of evidence about four cases, collected and organized to facilitate the analysis of the two frameworks. A comparative case study analysis characterizes these frameworks across multiple real contexts of use to facilitate their evaluation.

The findings identify opportunities to improve these frameworks and raise questions for further investigation into the design of such frameworks.

To facilitate adoption of the proposed approach, this report will describe the method in detail and document its application extensively. A set of appendices provides structured documentation to support the replication of the case study method in other contexts and in relation to other frameworks.

The report supplements and extends an article (DOI: 10.1002/asi.24209) published in the Journal of the Association for Information Science & Technology (JASIST). The report is structured as follows:

- **Background** explains the state of art in organizational assessment and maturity models in digital curation and the broader state of art on this subject in the wider fields of information systems and IT management.
- **Evaluated Frameworks** explains the organizational context of the research presented here. It then discusses and justifies the selection of frameworks chosen for evaluation.
- **An Assessment Procedure** describes a step-wise process to conduct an assessment of an organization.
- **Evaluation Method** describes the high-level methodology used for evaluating frameworks from analytic and empirical perspectives.
- **Process Evaluation** describes the evaluation of the frameworks’ design process.
- **Design Evaluation** performs a structured analytic evaluation of both frameworks.
- **Usage Evaluation** describes the empirical evaluation method in detail and reports on four assessment projects.
- **Analysis of Assessment Projects** summarizes the findings about the usage of frameworks.
- **Analysis of Frameworks** discusses characteristics of the evaluated frameworks.
- **Synthesis of Evaluation** explains empirically observed incidents using the results of the design evaluation.
- **Conclusions** summarizes findings and discusses the methods and its limitations.
- **Summary and Outlook** concludes the report and points to consequences and future work.
- **Appendices** provide the survey and interview questionnaires used in the case studies.
2 Background

This section provides an overview and history of organizational assessment in the digital curation field and prior work on the design of assessment frameworks. In doing so, it identifies the need for systematic evaluation and for a replicable evaluation method to support this.

2.1 Organizational Assessment in Digital Curation

Systematic assessments of an organization’s capability to curate digital resources over time can provide strategic direction for improvement, instill trust in stakeholders, and in some cases, help the organization to become certified according to established standards.

As a response, over the past decade, many models and tools have been created to facilitate and structure such assessments. The overall term “assessment framework” in this report refers to a variety of models, tools, and guidelines that can be used by practitioners in an organization to conduct an assessment of the current state of their digital curation abilities according to a set of criteria. These frameworks often use concepts such as maturity and capability to evaluate the ability of an organization to preserve and curate digital resources. An extensive review identified over a dozen frameworks and an increase in recent years in activities that develop new conceptual models to guide the focus and criteria of assessment (Maemura, Moles, & Becker, 2017). These include DRAMBORA, the Digital Repository Audit Method Based on Risk Assessment (McHugh, Ross, Ruusalepp, & Hofman, 2007; McHugh, Ross, Innocenti, Ruusalepp, & Hofman, 2008); DPCMM, the Digital Preservation Capability Maturity Model (Dollar & Ashley, 2015); and AIDA, the Assessing Institutional Digital Assets toolkit (Pinsent, 2009).

Assessment frameworks typically provide a model, i.e. a conceptual structure for diagnosing an organization. The framework’s terminology, perspective and structure determine what questions to ask about the organization and how to judge the evidence at hand to diagnose an organization. Since evidence plays a key role in establishing trust in digital curation (Beagrie et al., 2002; Ross & McHugh, 2006), the choices of which evidence to consider during the assessment project is an important aspect of the framework’s tools and guidelines. Similarly, how the assessment framework sets the bar for considering an activity “performed” or a goal “achieved” determines what constitutes an “established” or “mature” capability. For example, considering the activity of ingesting digital objects into a repository or archive, what makes Ingest an established capability? Is it when the digital repository software provides an operational interface for Ingest; or when the repository has performed its first Ingest operations on actual objects; or when the repository has assigned responsibility; or when it has regularly performed Ingest over a certain time of period? Or, is it established when a risk assessment has been conducted on the Ingest function?

The structure of an assessment framework and the focus it provides in the diagnosis are thus central for the outcomes. Assessment results in turn serve as vital input for strategic decisions about the future development of the organization with a digital curation mandate (McHugh, Ross, Ruusalepp, & Hofman, 2007; Dobratz, Rödig, Borghoff, Rätzke, & Schoger, 2010; Kulovits, 2013). This raises questions of how effectively the frameworks support practitioners in the assessment process and to what degree frameworks help to produce reliable insights that can guide future development.

In other words, there is a need to evaluate the framework’s quality in use - the “degree to which a product or system can be used by specific users to meet their needs to achieve specific goals with effectiveness, efficiency, freedom from risk and satisfaction in specific contexts of use” (ISO 25010:2011).

At present, little research exists on how effective different frameworks are in supporting an organization’s goals or facilitating improvement. Given the strategic importance of the assessment results, it is only natural to expect an assessment framework has been evaluated rigorously, and evidence about its usage is made available to potential adopters to enable them to evaluate the suitability
of the framework for their particular context (Jokela, Siponen, Hirasawa, & Early, 2006; Wendler, 2012).

It is essential to have empirical insights into the process of diagnosing and assessing an organization when developing such models (Hevner, March, Park, & Ram, 2004; McHugh et al., 2008) and empirical studies have contributed to the development of some frameworks, such as DRAMBORA (McHugh et al., 2008; Ross & McHugh, 2006). The insights needed for model development can only be obtained by studying models in use through systematically collecting and analyzing evidence of their functioning in practice.

However, a comprehensive review of publications about assessment in digital curation showed that despite the growing number of frameworks developed in recent years, most do not support their claims by disclosing evidence of their application, evaluation and validation (Maemura et al., 2017). While success stories of their application accompany some frameworks, none provide comprehensive evidence about characteristics or evaluation methods. The work presented here aims to address this gap.

Assessment frameworks come in different shapes, but share common traits. We follow Maemura et al. (2015; 2017) in the terminology used to distinguish common components. An assessment framework to support the process of systematic assessment contains at a minimum a structured model, i.e. a set of criteria grouped in a certain way that describe and structure the maturity and/or capability of an organization. The model is often accompanied by some documentation including a set of guidelines to support the process of conducting an assessment according to a specific method, and sometimes by a tool to support the assessment and/or produce a score. Common tools for assessment can include simple forms and templates or more complex software applications.

In this report, we use the term framework or assessment framework to refer to all of the components that together form a single approach, and the specific term model or maturity model to refer to the articulation of criteria and scoring mechanisms that are typically at the core of a framework. We use the term assessment for the scoring of organizational abilities and reserve the term evaluation to denote the characterization and study of frameworks, models and their use.

The available digital curation assessment frameworks follow very different approaches. Some focus on asserting compliance with such criteria catalogues as ISO 16363, others on self-assessment (McHugh et al., 2008). The digital curation community has also demonstrated avid interest in questions of self-assessment in recent workshops and meetings at IPRES 2013 (Becker & Cardoso, 2014), IPRES 2015 (Becker, Maemura, & Moles, 2015), IDCC 2016 (Lyon, Becker, Maemura, & Wei, 2016) and ASIST 2016 (Qin, Crowston, & Lyon, 2016). The discussions at these venues highlighted interest in medium-effort tools or frameworks to support organizational improvement and alternatives to time- and resource-intensive audits (Goethals, 2013; Becker & Cardoso, 2014). This corresponds to a long-standing interest in self-assessment in digital curation (McHugh et al., 2008).

As a lighter-weight alternative to formal processes of external audit or certification, frameworks that support reliable and systematic self-assessment can benefit organizations by supporting the diagnosis of the current state and facilitating the planning of specific steps towards improvement (McHugh et al., 2008; Pöppelbuß, Niehaves, Simons, & Becker, 2011). These self-assessment frameworks are of particular interest to us due to their potential value in supporting the iterative growth of an organization’s maturity.

### 2.2 The Design of Assessment Frameworks

Assessment frameworks often draw on concepts from the fields of process improvement and IT management. In particular, the assessment of an organization is often framed in terms of one or more capabilities, where a capability refers to the sustained ability of an organization to achieve an identified goal through a combination of people, technology and process (Pöppelbuß et al., 2011); or maturity, i.e. the position of an organization on an evolutionary path with respect to an aspect of interest, such as a specific capability (Mettler, Rohner, & Winter, 2010).
The growing interest in assessment frameworks in digital curation mirrors a similar growth of interest in maturity models, a type of framework that has been adopted in many other domains (Wendler, 2012). Maturity models are the most common type of conceptual structure used for assessment in that they describe progressive stages through which organizations are expected to progress over time as they are improving and maturing. They can cover the maturity of specific capabilities defined through processes and activities performed in the organization, but can also focus on other aspects of maturity such as awareness, communication, or accountability. The specific meaning of capability and maturity are often not explicitly defined, but implicit in the articulation of criteria and dimensions.

Maturity models have been developed for a range of fields related to the management, development and deployment of IT-supported organizations. A wealth of theoretical and empirical research on models for assessment and improvement exists (Jokela et al., 2006; Mettler et al., 2010; Wendler, 2012; Maier, Moultrie & Clarkson, 2012; Pöppelbuß et al., 2011). A very common structure for such models is the maturity grid (Maier et al., 2012). The literature in these fields often uses the term “maturity model” to refer to the entire framework, including the model, guidelines, and tools.

Capability maturity models are a specific type of maturity models that focus on a set of capabilities for a specific domain, such as digital preservation or software development. The well-known Capability Maturity Model (CMM) for software engineering developed by the Software Engineering Institute (SEI) is commonly seen as the most influential and highly developed capability maturity model. The original CMM for software engineering processes divides capabilities needed for software development into 18 key process areas with specific activities and maps each capability to five maturity levels: Initial (Level 1), Repeatable (2), Defined (3), Managed (4), and Optimizing (5) (Paulk, Curtis, Chrissis & Weber, 1993). The SEI subsequently developed a comprehensive and rigorous framework for applying the CMM, which outlines very specific requirements for different methods of assessment with increasing degrees of rigor and reliability (SCAMPI Upgrade Team, 2011). Reliability in this context is the degree to which independent applications will arrive at the same assessment of a given organization. Many other models and frameworks have been derived from and created following this work.

2.3 Evaluating Assessment Frameworks

In a survey of assessment frameworks for digital curation, Maemura et al. (2015) identified 14 available frameworks and grouped them by the primary purpose or scenario for their use (initial planning, improvement, and certification), as well as their modes of application (self-assessment or third-party assisted) and supporting tools, guidance, and evidence. A subsequent systematic review and mapping of the literature on organizational assessment in digital curation determined the extent and type of publications on assessment that exist in the field (Maemura et al., 2016).

The analysis mapped each publication to one of four categories: model development (developing and presenting a new model for assessment), model application (describing and presenting the assessment results of a real organization), model validation (testing a model’s accuracy, reliability, and effectiveness), and meta-theoretical reflections on the purposes of or approaches to assessment. Of the 80 papers found through the literature search, most focused on development and application, with only two papers addressing model validation and nine papers on meta-theoretical reflections. The conclusions showed that the development of models is an increasingly common research activity, but often overlooks the provision of specific methods and step-wise guidance for carrying out the assessment. Applications of models in real world settings are often reported uncritically as ‘success stories’ with little evaluation or supporting evidence. Still, a number of third-party applications have been reported. For example, Maemura et al. (2016) identified seven such reports for DRAMBORA, six for the Data Seal of Approval (Harmsen, 2008), and one for AIDA, and summarize the conclusions drawn in these reports.

While many such models and tools have been created for assessing organizations, these tools have not undergone a level of scrutiny in their development comparable to that expected from organizations themselves (Maemura et al., 2016). This is not unique to digital curation: maturity models are often
proposed without rigorous evaluation, or with evaluation in the form of case studies designed to confirm the value of the model rather than evaluate and critically compare it to existing models (Wendler, 2012). The central role of these models in determining strategic decisions suggests that this possible confirmation bias masks the identification of shortcomings in the models, and that more rigorous research is needed to critically evaluate and validate the models for assessment and inform the development of new models or further design iterations of existing models (Maier et al., 2012; Pöppelbuß et al., 2011).

Despite the importance of assessments for organizations, little is often known about the assessment process and the outcomes of these cases. Understandably, many organizations have concerns about sharing the results from their assessments publicly (McHugh et al., 2008; Ross & McHugh, 2006), and many of the case studies that have been published were conducted by the creators of frameworks for promotion and to demonstrate their use (Wendler, 2012). A number of reports on external audits are available from organizations such as the Center for Research Libraries, but do not focus on a detailed description of the assessment process or an evaluation of the framework used for assessment. Similarly, several reports of self-assessments with ISO16363 have been published, but do not present systematic evidence that can be used for independent analysis or comparison. This growth of shared evidence is a core concern of the further development of the digital curation field (NDSA, 2014). In the case of assessments, this should thus include a structured description of the assessment process and its context and outcomes, and the role of the assessment models in the assessment.

3 A Method for Evaluating Assessment Frameworks

Our method for evaluating assessment frameworks encompasses three complementary perspectives – process evaluation, design evaluation, and usage evaluation. Figure 1 illustrates the conceptual relationships between the intangible procedures used to design a framework and to apply it in a given context, the tangible documents and tools we can review and analyze, and the evidence gathered in empirical studies of assessment frameworks used in assessment projects.

The diversity of these three lenses requires an approach capable of including the analytical methods required to examine both a framework’s design process and subsequent product and the empirical methods needed to evaluate quality-in-use. For this reason, our method combines analyses based on Design Science Research (DSR) theory with empirical evaluation based on a multi-case study approach where the assessment frameworks are used by practitioners. DSR is a well-established paradigm in Information Systems (Gregor & Hevner, 2013; March & Smith, 1995) that focuses on the systematic design and evaluation of models and tools, specifically in the context of information systems and IT management. It builds on Herbert Simon’s argument that designed artifacts can and should be studied as rigorously as natural phenomena (Simon, 1969) and strives to create research that is meticulously conducted, well documented, and contributes to the knowledge bases of both researchers and practitioners (Hevner, March, Park, & Ram, 2004; March & Smith, 1995). DSR contributes principles, methods and guidelines to the design process that work iteratively to create designed artifacts and evaluate them with clear requirements applied through empirical or analytical testing. In information systems research, some methodologies and procedures for design science (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), have been highly influential (Gregor & Hevner, 2013; Hevner & Chatterjee, 2010). Within this paradigm, assessment frameworks constitute designed artifacts used for specific purposes in particular contexts.

**Process Evaluation:** The framework’s design process is evaluated based on the eight “Requirements for the development of maturity models” defined by J. Becker, Knackstedt and Pöppelbuß (2009). For each framework under investigation, we evaluate its adherence to each of their requirements. To do this, we rely on publicly available documentation provided by the creators of each model to make cautious inferences about the design process.

**Design Evaluation:** The framework itself is evaluated by studying the model using the “Design Principles for Maturity Models” developed by Pöppelbuß and Röglinger (2011). For each framework under investigation, we evaluate its adherence to each design principle. To do this, we review all available documentation, including supporting material such as glossaries, tools, and documents that describe the framework’s procedures.

**Usage Evaluation:** Evaluating how a given framework works in a given context of application, and why, implies a number of commitments.
1. The evaluation must be grounded in empirical studies of the practice of assessment to ensure practical relevance (Hevner et al., 2004). In doing so, it must study assessment in its real-world context, not an artificial setting as is done in experiments.

2. Evaluation must address the interest of the practitioners in obtaining relevant insights into their organization, and simultaneously provide a frame that studies the practice of assessment to better understand it.

3. Beyond recording what happened, the evaluation method should enable researchers to engage in explanation building to provide insights about the frameworks applied in the assessment project that have relevance to practitioners, framework developers, and researchers.

Usage evaluation is centered on the assessment project, the complete set of activities undertaken in the process of assessing an organization. The assessment project produces the scores, rankings, outputs, and reports that constitute the assessment results and indicate the organization’s degree of maturity. While these results are the motivation of the organization to perform the assessment, we focus here on the observations and perceptions of the project and what it reveals about the framework in use. Case summaries captured the insights we obtained from observing the project along with our reflections and those of the assessors.

The method of choice for studying such bounded contemporary social phenomena in their real context is case study research. It uses multiple sources of evidence to develop a structured understanding of the phenomenon in question and is especially applicable when questions of ‘how’ and ‘why’ are posed and the investigator has little control over events (Eisenhardt, 1989; Yin, 2014). As such, case studies have been a valuable instrument for researchers in digital curation (Kansa, Kansa, & Arbuckle, 2014; Pryor, 2009, 2013; Wallis, Borgman, Mayernik, & Pepe, 2008). For example, case studies were instrumental in developing the DRAMBORA risk assessment framework (McHugh et al., 2008). Guidance on how to select, conduct, analyze and report case studies can be helpful in enabling more case studies to be conducted and shared as a basis for discussion (Yin, 2011).

The framework’s quality in use is thus evaluated by conducting assessment projects in partnership with organizations who perform organizational assessment. The assessments are carried out by practitioners to produce meaningful results. A set of data collection methods is employed to gather evidence about the assessment project that includes observations of the process of assessments, a review of the organization’s perspective before the assessment, and a semi-structured interview with the practitioners about the value and use of the assessment. This will be described in detail in the section Usage Evaluation.

In our application of the method, we increase the robustness of the resulting empirical insights by using a multi-case study research design inspired by Robert K. Yin (2011; 2014). In the multi-case study design, frameworks are evaluated across multiple organizations, with each combination of framework and organization resulting in one assessment project or “case.” To do this, the assessment procedure explained above is used to conduct the assessments, and then each case is evaluated separately and comparatively. This method is described in detail in Usage Evaluation.

In the subsequent sections, we provide an in-depth look into each evaluation perspective, accompanied by an actual evaluation of DPCMM and AIDA based on those perspectives. In the section Process Evaluation, we will delve into Becker, Knackstedt and Pöppelbuß’s 8 requirements for maturity models and how we evaluated the frameworks based on those requirements. This is followed by our evaluations of DPCMM’s and AIDA’s design processes. In the section Design Evaluation, we will cover Pöppelbuß and Röglinger’s three sets of design principles (2011) and how, or if, each assessment framework meets them. Finally, in the section Usage Evaluation, we provide a description of our multi-case study method followed by a usage evaluation that includes four case studies of assessment projects (DPCMM in use at ASA, DPCMM in use at UTL, AIDA in use at ASA, and AIDA in use at UTL). The subsequent sections will analyze the evidence collected through these cases to study in detail the assessment projects (i.e. the use of frameworks) and the frameworks themselves in use.
4 The Selection of Frameworks for Evaluation

4.1 Organizational Context

This evaluation report is the result of a collaboration of the research group with practitioners in two organizations that provided the context and setting for our empirical evaluation of assessment projects.

The first organization is the Austrian State Archives (ASA), the national archives of the Republic of Austria. It has a legal mandate to collect, preserve, and make available all archival holdings of the federal government for research, legal and administrative purposes.

At ASA, the self-assessments were conducted by the head of the Digital Archive division, and focused on the digital archive and its holdings, including electronic records created by the federal government. A goal of undertaking self-assessments was to support improvement with the goal of achieving Trusted Digital Repository status in the future.

The second is the University of Toronto Libraries (UTL), the academic library system of the University of Toronto. Their digital collections include over one million resources managed by UTL’s Information Technology Services department. At UTL, the department’s Digital Preservation Librarian conducted the self-assessments. The assessments focused on the Fedora-based repository and TSpace, an open access repository for theses and dissertations. A goal of undertaking self-assessments was to evaluate the current state of UTL’s digital preservation program and provide the findings to the newly established Digital Preservation Policy Working Group.

Both organizations have well-established digital preservation programs, while each of the participants in our research has extensive experience and expertise in the field, including prior experience with a full ISO 16363 audit (UTL) and an ISO 16363 self-assessment (ASA). This influenced our selection of frameworks, and has implications on our conduct and findings:

1. We expected little uncertainty to arise during the assessment projects about the understanding of the organization’s abilities.
2. We expected little need for the study of detailed explanatory materials about digital curation frameworks or terminology.
3. Our findings about the individual frameworks cannot easily be assumed to hold in organizational contexts where the assessors have little experience or the organization has very low capabilities.

We will return to these implications in the discussion section.

4.2 Selection Criteria and Frameworks

The frameworks evaluated here were selected from those surveyed by Maemura et al. (2017), based on the criteria described and justified in Table 1. These criteria were developed to make the study relevant and valuable for both practitioners and researchers.

Table 1 Selection criteria for the frameworks studied

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>Focus on self-assessment, independent of third-party support, not external certification.</td>
<td>Both organizations involved in the study had no interest in external certification at the point of study, but a strong interest in an internal diagnosis that provided meaningful insights.</td>
</tr>
</tbody>
</table>
The scope of the model should coincide with concerns of digital preservation in the participating organizations. The assessments were meant to provide meaningful insights to the organizations about their digital preservation capabilities.

Both participants have concerns about the restrictive nature of compliance frameworks and were interested in concepts of improvement over time and maturity. This coincided with the researchers’ interest focus.

It must be feasible (1) to complete the assessment within the organization and (2) to study the process and its results through multiple sources of evidence. This includes the availability of documentation of the model, method, and tools.

The scope excluded general frameworks for assessing IT Governance such as COBIT (IT Governance Institute, 2007) and narrow frameworks for assessing particular aspects of digital curation or digital preservation. We selected two frameworks that met all five criteria and were particularly interesting to both the organizations and researchers. The frameworks were suggested by the researchers after consulting with the assessors, and approved by the practitioners.

1. **The Digital Preservation Capability Maturity Model (DPCMM)** draws heavily from the OAIS Reference Model (ISO 14721:2012), TRAC (CRL & OCLC, 2007), and ISO 16363 (2012). The framework has been applied at a number of institutions in the United States and Canada, and the creators continue to refine the DPCMM and apply it in their ongoing consulting work (Dollar & Ashley, 2014; Dollar, Ashley & Misic, 2014). The assessment projects here used the DigitalOK online tool (http://digitalok.org/).

   The DPCMM was most interesting for the researchers for its clear positioning as a Capability Maturity Model, and for the ASA, as a low-effort model that claims to provide a comprehensive assessment and alignment with the OAIS.

2. **The Assessing Institutional Digital Assets (AIDA) toolkit** was developed to assess the digital asset management capability of UK higher education organizations, following the conceptual framework of the “Five Organizational Stages of Digital Preservation” and the “three-legged stool” of Organization, Technology and Resources (Kenney & McGovern, 2003). It has been applied at UK universities, as well as in the United States (Pinsent, 2009; Miller, Blake & Sorsby, 2012).

   AIDA was most interesting for the researchers for its flexible approach and its model of organizational change, and for UTL, as a model that arose from the education sector.

The next sections provide a brief overview of each of these frameworks, focusing on the structure of the models.

### 4.3 DPCMM: The Digital Preservation Capability Maturity Model

The DPCMM framework consists of the model (documented in the latest version 2.7, released in July 2015) and the online tool “DigitalOK.” The model consists of 15 dimensions or components. Eight
components assess “Digital Preservation Infrastructure” and seven components assess “Digital Preservation Services,” as illustrated in Figure 2.

![Figure 2 DPCMM components (redrawn from Fig. 1 (p9) and Fig. 2 (p12) in Dollar & Ashley, 2015)](image)

For each component, the model defines five levels representing “Stages of Digital Preservation Capability Maturity” (Dollar & Ashley, 2015, Fig. 1, p. 3). Each component is scored and the scores for each are aggregated to a Composite Index Score. Higher stages are characterized by a decreasing amount of records being at risk. The model identifies a threshold for conformance to ISO 14721 (OAIS, 2012) through its stages, which are described in Table 2.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, Nominal</td>
<td>Most, if not all, electronic records meriting long-term preservation are at risk.</td>
</tr>
<tr>
<td>2, Minimal</td>
<td>Many electronic records meriting long-term preservation are at risk.</td>
</tr>
<tr>
<td>3, Intermediate</td>
<td>Some electronic records meriting long-term preservation remain at risk.</td>
</tr>
<tr>
<td>4, Advanced</td>
<td>Few electronic records meriting long-term preservation are at risk.</td>
</tr>
<tr>
<td>5, Optimal</td>
<td>No electronic records meriting long-term preservation are at risk.</td>
</tr>
</tbody>
</table>

The DigitalOK tool, which accompanies the DPCMM, supports the assessment through a series of online forms. First, the tool presents a survey to gather background information about the organization and assessor. Then, a form for each of the 15 components is presented. These forms list a series of statements and the assessor determines if each statement represents the situation of their organization. The statements are presented individually and are not explicitly linked to one of the five levels. After completing each form, an aggregate score is generated and the scores for each component are presented in a final PDF report the assessor can download.
4.4 AIDA: Assessing Institutional Digital Assets

The AIDA Toolkit consists of a handbook, a project report, a scorecard template, and a weighted scorecard spreadsheet. The model is built on the metaphor of the Three-Legged Stool of Organization, Technology and Resources, and stages that conceptualizes digital preservation through increased awareness, maturity and externalization (Kenney & McGovern, 2003). The maturity levels are: (1) Acknowledge, (2) Act, (3) Consolidate, (4) Institutionalize, and (5) Externalize. AIDA operationalizes these stages across three areas through 31 areas or dimensions, as illustrated in Figure 3.

For each leg, a scoring sheet (called “Assessment Scorecard”) is provided that includes some background information and then, for each stage, a description of typical characteristics of organizations on that level in the particular dimension. These stage descriptions are provided for both the department and the institution as a whole to support assessment of organizational units in larger organizations and facilitate comparison.

![Figure 3 Overview of the AIDA model](image)

Since the scorecards do not facilitate the documentation of justification of scoring, the research team created a simplified template based on the scorecard to document the assessment. It consists of a table with a row for each dimension (e.g. Organization Leg, Mission Statement, and Departmental Level) which contain an identifier and short descriptor, a column to capture the stage number (1-5), and a column to write comments and rationale.

The models target different domains and types of organization: DPCMM is oriented towards compliance with OAIS and the preservation of electronic records in archives, but claims applicability to libraries and museums. AIDA is aimed at the management of digital assets in higher education institutions. Contrasting these approaches allows for critical reflection on the nature of maturity in this field and how assessors perceive the usefulness of different conceptions, but also provides a perspective on the flexibility and adaptability of the frameworks across organizational environments. Table 3 summarizes the key characteristics of DPCMM and AIDA.

---

<table>
<thead>
<tr>
<th>Name</th>
<th>Digital Preservation Capability Maturity Model © (DPCMM)</th>
<th>The Assessing Institutional Digital Assets self-assessment toolkit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acronym</td>
<td>DPCMM</td>
<td>AIDA</td>
</tr>
<tr>
<td>Primary Source</td>
<td>Background and Performance Metrics Version 2.7 (Dollar &amp; Ashley, 2015)</td>
<td>Toolkit Mark II (ULCC, 2009a)</td>
</tr>
<tr>
<td>Addressed topic</td>
<td>Digital preservation</td>
<td>Digital asset management</td>
</tr>
<tr>
<td>Origin (academic vs. practice)</td>
<td>Consultant team</td>
<td>Practitioner-researcher team</td>
</tr>
<tr>
<td>Audience</td>
<td>Internal, management-focused, business and technology</td>
<td>Internal, management-focused, business and technology</td>
</tr>
<tr>
<td>Year of publication</td>
<td>V2.7 - July 2015</td>
<td>Mark II - February 2009</td>
</tr>
<tr>
<td>Access</td>
<td>Model and tool are freely available, though built around consulting service</td>
<td>Freely available, but online tool is no longer maintained.</td>
</tr>
</tbody>
</table>
5 An Assessment Procedure

Many assessment frameworks do not specify an explicit procedure for assessment (Maemura et al., 2016; Frick, Küttner, & Schubert, 2013).

In this absence, we developed a general procedure for assessing a framework to ensure consistency. This method is broadly based on the steps defined by the appraisal methodology of CMMI (Hayes et al., 2005; SCAMPI Upgrade Team, 2011) but, due to its focus on self-assessment, omits the rigorous collection of ‘objective evidence’ that those procedures require. The procedure structures the assessment project into four stages, as illustrated in Figure 4.

1. **Phase 1: Plan and prepare** - Information about the organization is documented and the assessment’s scope is defined. It determines what specific organization, department, repository, or set thereof is the focus of the assessment. This phase uses the tools provided by each framework for documenting the relevant information about the organization, the assessor, and the scope.

2. **Phase 2: Conduct assessment** - The organization’s assessor progresses through the criteria of the model in sequence. For each component of the assessment model, the assessor determines which level most accurately captures the capabilities of the organization. The result is recorded with the tools provided.

3. **Phase 3: Review assessment** - Before proceeding to the aggregated results, the organization’s assessor reviews all answers and resulting scores and may reconsider and change initial answers or provide additional notes and rationale.

4. **Phase 4: Review results** - The individual components are aggregated and presented as a final assessment result, through the tools provided in the framework. The organization’s assessor reviews this final presentation or scoring, and any further information provided by the framework to characterize or aid in the interpretation of the result.

---

**Figure 4 A basic assessment procedure to structure an assessment project**
6 Process Evaluation

Table 4 presents the requirements developed by Becker, Knackstedt, and Pöppelbuß for maturity model development (2009) alongside how they can be evaluated. This is followed by the results of our evaluation of the design process of DPCMM and AIDA.

Table 4 Methods for evaluating requirements for maturity model development

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Method for evaluating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R1 - Comparison with existing maturity models:</strong> “The need for the development of a new maturity model must be substantiated by a comparison with existing models. The new model may also just be an improvement of an already existing one.”</td>
<td>This is inferred through statements describing how relevant predecessors are insufficient and why a new model is necessary for this domain.</td>
</tr>
<tr>
<td><strong>R2 - Iterative Procedure:</strong> “Maturity models must be developed iteratively, i.e., step by step.”</td>
<td>This can be inferred from descriptions of the steps followed and changes made to the model through design iterations. The existence of multiple versions of a model is another indication, but does not in itself demonstrate systematic iteration.</td>
</tr>
<tr>
<td><strong>R3 - Evaluation:</strong> “All principles and premises for the development of a maturity model, as well as usefulness, quality and effectiveness of the artifact, must be evaluated iteratively.”</td>
<td>This can be inferred from full reports of testing/studies that report on its usefulness, quality, and effectiveness, or (less strongly) through statements made in other documentation about testing that was conducted during development.</td>
</tr>
<tr>
<td><strong>R4 - Multi-methodological Procedure:</strong> “The development of maturity models employs a variety of research methods, the use of which needs to be well-founded and finely attuned.”</td>
<td>Demonstrating the fulfilment of this requirement will only be possible if sufficient evidence is provided by the creators about the design process and methods used in developing their model (see R8).</td>
</tr>
<tr>
<td><strong>R5 Identification of Problem Relevance:</strong> “The relevance of the problem solution proposed by the projected maturity model for researchers and/or practitioners must be demonstrated.”</td>
<td>Demonstration of the relevance of the problem to the community can be achieved through citing literature that calls for assessment in the domain, or by other means that identify demand in practice.</td>
</tr>
<tr>
<td><strong>R6 Problem Definition:</strong> “The prospective application domain of the maturity model, as well as the conditions for its application and the intended benefits, must be determined prior to design.”</td>
<td>This is difficult to determine after the fact for frameworks not developed as part of a clearly scoped project, since conditions and intended benefits may be adapted post hoc.</td>
</tr>
<tr>
<td><strong>R7 Targeted Presentation of Results:</strong> “The presentation of the maturity model must be targeted with regard to the”</td>
<td>For the digital curation community, this requirement can be met through publicly accessible project reports as well as</td>
</tr>
</tbody>
</table>
conditions of its application and the needs of its users.”

presentations in community venues such as conferences.

**R8 Scientific Documentation:** “The design process of the maturity model needs to be documented in detail, considering each step of the process, the parties involved, the applied methods, and the results.”

We considered each element here separately and attempted to find documentation of each of: the overall process; each step in the process; the parties involved in the design; the methods; and the results. Since DSR methodology was not followed by the developers of the models studied here, the rigour expected of scientific documentation is often absent.

### 6.1 Process Evaluation for DPCMM

The results of the process evaluation, summarized in Table 5, reflect that DSR methodology was not followed in the design process for DPCMM. There is only a partial review of existing models, no evaluation, and no thorough explanation or documentation of design procedure. However, the creators of DPCMM do identify the problem their framework is attempting to address, and they present the framework itself and how to apply it in practice.

**Table 5 Evaluation of the design processes of DPCMM**

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Fulfilment</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1 - Comparison with existing models</td>
<td>Partial</td>
<td>DPCMM makes reference to the original CMM and other assessment frameworks and explains the aim of the DPCMM in relation to OAIS and ISO 16363, but provides no thorough review of existing models.</td>
</tr>
<tr>
<td>R2 - Iterative Procedure</td>
<td>Partial</td>
<td>Revisions were made to DPCMM and a new version was released following multiple applications, but no explanation is given of the steps followed in this process, nor how, or if case evaluation influenced those changes. Also, the latest model version does not match statements in the current tool.</td>
</tr>
<tr>
<td>R3 - Evaluation</td>
<td>No evidence</td>
<td>Documentation does not address evaluation of principles and premises for the development of the model, or its usefulness, quality and effectiveness.</td>
</tr>
<tr>
<td>R4 - Multi-methodological Procedure</td>
<td>No evidence</td>
<td>No documentation of development procedure to demonstrate a multi-methodological approach.</td>
</tr>
<tr>
<td>R5 – Identification of Problem Relevance</td>
<td>Yes</td>
<td>Relevance is noted, and an assessment commissioned by the Council of State Archivists demonstrates relevance.</td>
</tr>
<tr>
<td>R6 - Problem Definition</td>
<td>No evidence</td>
<td>No documentation of the early design phases is available to judge initial design premises.</td>
</tr>
</tbody>
</table>
6.2 Process Evaluation for AIDA

Here, the results differ somewhat from DPCMM and show that the methodology followed in the design process for AIDA was at least partially consistent with DSR guidelines. A review of existing models was done as part of the development process, the framework design was iterative (though those iterations were not explained), and pilot studies were conducted. A clear problem statement was defined, and the design products include an in-depth project report on the aims and applications of the framework alongside partial documentation of the design process. Table 6 summarizes the process evaluation for AIDA:

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Fulfilment</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1 - Comparison with existing maturity models</td>
<td>Yes</td>
<td>A first step in the development process analyzed LIFE, DAF, DAAT and DRAMBORA. The need for a new model is noted to differentiate AIDA from its predecessors.</td>
</tr>
<tr>
<td>R2 - Iterative Procedure</td>
<td>Partial</td>
<td>There is evidence of iteration, including changes made after the initial case studies and the release of two versions of the Toolkit, but there is no explanation about the iterative steps followed within the design process.</td>
</tr>
<tr>
<td>R3 - Evaluation</td>
<td>Partial</td>
<td>Pilot studies demonstrate some evaluation reflecting on assumptions and participants’ judgment of usefulness. Framework quality, effectiveness, and design process are not evaluated.</td>
</tr>
<tr>
<td>R4 - Multi-methodological Procedure</td>
<td>Partial</td>
<td>Some evidence including a review of existing models, a model crosswalk, prototyping, and case studies.</td>
</tr>
<tr>
<td>R5 – Identification of Problem Relevance</td>
<td>Yes</td>
<td>Not directly addressed, but support received through</td>
</tr>
<tr>
<td>R6 - Problem Definition</td>
<td>Partial</td>
<td>The overall application domain and intended benefits were identified initially, but the scope and conditions for application were not.</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>R7 – Targeted Presentation of Results</td>
<td>Yes</td>
<td>Project reports and presentations described the aims and applications of the framework.</td>
</tr>
<tr>
<td>R8 - Documentation</td>
<td>Partial</td>
<td>Overall steps and parties involved in design are noted, but a traceable account of each step (method and results) is not given.</td>
</tr>
</tbody>
</table>
7 Design Evaluation

The valuation of the frameworks’ design is structured by the three sets of design principles suggested by Pöppelbuß and Röglinger (2011), following the approach of Maier et al. (2012). For each model, we discuss basic, descriptive and prescriptive aspects guided by the criteria, provide a rationale for the evaluation, and summarize the findings in a table.

7.1 Design Evaluation for the DPCMM

In terms of basic design principles, DPCMM provides basic information (1.1) for the intended use, although it omits explicit prerequisites for applicability (1.1a) and would benefit from more clarity in this regard. It defines its target group (1.1c) very widely in stating it aims to be applicable to “archives, library, records management and business communities” as well as organizations in both the public and private sectors (Dollar & Ashley, 2015, p. 2). The class of entities under investigation (1.1d) is not explicitly defined. The model states a focus on capabilities, but the intended definition of capability is unclear. It addresses both specific technological system components and socio-technical services and processes, without explaining the rationale or connections between them. It does not characterize or explain the relationships between process maturity, systems capability and growth over time.

Concepts of Maturity (1.2) are not defined, which is common: As Pöppelbuß & Röglinger (2011) note, “most maturity models do not define but circumscribe maturity” (p. 6). The DPCMM claims to be a CMM, but deviates significantly from the CMM core concepts and structure. The lack of clarity in theoretical foundations is reflected in how maturity levels and maturation are defined. The model does not follow the structure or intentions of the SEI CMM: While the CMM defines key process areas (KPAs), each with a set of activities and goals, the DPCMM equates components to KPAs without considering their specific structure. Levels 4 and 5 in the CMM and CMMI intentionally do not introduce domain-specific KPAs but focus on cross-domain activities of measuring, improvement and learning. This view of maturity is not reflected in the DPCMM, where criteria for levels 4 and 5 routinely describe new kinds of activities and metrics. The DPCMM’s scoring of individual criteria and cumulative ‘index score’ provided in the results are not compliant with the Appraisal Requirements for the CMM (SCAMPI Upgrade Team, 2011). The latter also specify that scores cannot be provided unless a rigorous evaluation method is followed to link each criterion to evidence.

To structure its concept of maturity in the domain, the DPCMM claim to build on the OAIS model. Some of the 15 dimensions are derived directly from OAIS Functional Entities such as Ingest, Archival Storage, and Access, while the origins of others are unclear. More problematically, there is no thorough consideration of maturity in DPCMM, and its maturity levels focus on increased capability rather than increasing maturity. This reduces it in fact to a capability model, where the capability - a functional concept - claims to be aligned with the OAIS. The thresholds of horizons between maturity levels deviates significantly from the CMM, on which the model claims to build. Increasing levels of maturity are associated with the concept of fewer and fewer electronic records being at risk. However, the assumption that the increasing level of maturity translates into a reduced number of records that are at risk in an organization is not founded in domain literature and seems questionable: Depending on the nature of gaps in capability and maturity, all records could be at uniform risk, for example when no sustainability strategy has been developed. The DPCMM presents the highest level of maturity as “Optimal: In Stage 5 no electronic records that merit long-term preservation are at risk” (Dollar & Ashley, 2015, p. 9). In contrast, the highest level of the CMM and most other maturity models is characterized as “optimizing” not “optimal” - where optimizing indicates continuous learning. By suggesting a stage is optimal, the DPCMM risks delivering a false sense of ultimate achievement.

Problematic conceptions of maturity are also evident in the lower levels. For example, “Digital Preservation Strategy” (component 2) describes Level 1 (Minimal) as “The strategy calls for accepting electronic records in native formats on an ad hoc basis and keeping the bit streams alive until software and other resources are available to transform the records into open standard technology neutral file formats.” (Dollar & Ashley, 2015, p. 17). While minimal, this strategy can be a cost-effective and
resource-efficient strategy for many situations (Rosenthal, 2010; Jurik et al., 2015). More critically, Level 3 - within OAIS conformance - requires that the “strategy calls for transformation of electronic records in five (5) selected native file formats to preferred preservation formats at ingest…” This can be potentially destructive and contrary to current understanding of migration procedures and OAIS requirements. The number five has no grounding and little meaning without organizational context. Level 4, similarly, requires mandatory transformation upon ingest. For many organizations, this will be counterproductive, since automatic transformation into newer forms is not commonly provably correct and can harm the authenticity of records. The highest level of strategy mandates conversion without addressing the need for assessment and preservation planning.

Similar concerns arise in other areas. In the Designated Community metric, the distinction between levels is entirely determined by the degree of documentation that exists. In leading maturity models in relevant areas such as IT Governance (IT Governance Institute, 2007), the degree of documentation is commonly only one of several dimensions and not the ultimate determining factor of maturity. This multi-dimensional view has been applied to Digital Preservation maturity as well (Becker, Antunes, Barateiro, Vieira, & Borbinha, 2011).

Similar to the class of entity under investigation, the underpinning theoretical foundations with respect to evolution and change (1.2d) of these entities are left underspecified. However, the DPCMM anticipates that an organization will develop the functional components to a degree where no records are at risk, but leaves open how this evolution takes place.

DPCMM uses domain-specific concepts (1.3) adopted from the OAIS reference model. However, these concepts and terms are incomplete and not consistently applied, and the Glossary of Terms provided in the model does not fully attribute its sources.

Documentation for the model was found to be appropriate for the target group of users, with conference presentations and openly available publications targeted at the domain.

Descriptive use requires the ability to consistently apply a model to diagnose an organization. Intersubjectively verifiable criteria (2.1) require consistency of the structure and language used in individual criteria. For DPCMM, criteria are phrased in specific terms, but inconsistencies in the terminology suggest that the criteria may be judged differently by different assessors. Many criteria are not atomic and consist of multiple individual assertions linked with the word “and”. This makes it impossible to apply them to precisely describe the possible states of an organization’s activities.

The model does not provide a comprehensive procedure as part of the Assessment methodology (2.2). DPCMM provides no methodology, advice on assessment criteria or advice on adaptation and configuration of the criteria, nor documents that provide expert knowledge from previous applications. An implicit method for assessment is however embodied in the DigitalOK tool, which structures the procedure through the affordances of its user interface.

In terms of prescriptive use, DPCMM provides no specific improvement measures (3.1), no decision calculus for selecting improvement measures (3.2), nor a methodology for making these decisions (3.3).

The findings are summarized in Table 7.

<table>
<thead>
<tr>
<th>Principle</th>
<th>Evaluation of the DPCMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Basic information</td>
<td>No prerequisites for applicability are noted. Adherence to the OAIS Reference Model is implied but not stated. <strong>Target group</strong> is defined very broadly. <strong>Class of entities under investigation</strong> is not defined. Components address capabilities of both technological systems and socio-technical services and processes.</td>
</tr>
</tbody>
</table>
7.2 Design Evaluation of AIDA

In terms of basic design principles, AIDA provides basic information (1.1) for the intended use but omits explicit prerequisites for applicability (1.1a) and would benefit from more clarity in this regard. It addresses its target group (1.1c) in the higher education sector, but does not provide much clarity in defining the class of entities under investigation (1.1d). The framework distinguishes between the department and institution, as well as the three legs of organization, technology, and resources, but does not differentiate between capability, processes, systems, or other possible elements. The elements within the three legs vary in that they investigate a wide range of processes, systems, and resources such as funding availability, backup facilities or staff training programs, but this diversity is not explained or addressed.

Concepts of Maturity (1.2) are not clearly defined, but AIDA’s maturity levels are based on the Kenney & McGovern Five Organizational Stages of Digital Preservation: (1) Acknowledge, (2) Act, (3) Consolidate, (4) Institutionalize, (5) Externalize. The assumption is that as an organization matures, its digital preservation services are increasingly developed and institutionalized, but also increasingly externalized until they “embrac[e] inter-institutional collaboration and dependency” in their highest organizational stage (Kenney & McGovern, 2003). Granularity is provided by AIDA’s distinction between the three legs, each further divided into components, and between departmental and institutional maturity levels for each component. However, it is unclear how the overall assessment through a weighted scorecard reflects the maturity of each sub-component.

Since AIDA does not define the entities under investigation, the underpinning theoretical foundations with respect to evolution and change (1.2d) of these entities are left underspecified. However, AIDA builds on the three dimensions of Organization, Technology and Resources, which emphasizes the importance of a balanced and coordinated progression on all three, and the evolution is meant to progress towards externalization of services according to the Stages of Growth.

AIDA provides informal descriptions of domain-specific concepts (1.3) such as digital assets and digital asset management. Each of the dimensions is related to at least one of four separate sources (ULCC, 2009b), but the nature of the illustrative statements used for assessment is reflected in a broad use of terms.
Documentation was found to be appropriate for the target group of users and publicly available. **Descriptive Use** requires consistency of the structure and language used in individual criteria.

In terms of *Intersubjectively verifiable criteria* (2.1), some criteria clearly identify evidence such as a Mission Statement or Business plan, while other criteria rely on subjective opinions such as the degree of regularity with which a process takes place or the level of awareness that exists across the department or institution. Thresholds between the stages are characterized through exemplars rather than clearly defined. This conscious design decision is meant to allow the assessor to interpret how each level might be conceived within their own organization, and the illustrative examples are meant to lower the entrance barrier to application. The lack of explicit criteria suggests that the model cannot be applied consistently across assessments (for a descriptive use) nor provide accurate guidance for improvements based on specific maturation paths. It remains to be evaluated what effects this has in practice.

AIDA does not provide a comprehensive procedure as part of the *Assessment methodology* (2.2), but includes a simple procedure model in the “How do I use this toolkit” section. Some advice on the assessment criteria is provided for each element, as well as illustrative examples (and how they might be adapted to specific cases). Explanatory notes accompany some elements. There is encouragement to adapt the model as necessary for contexts where it might not fit, but no specific guidance on how or when adapting the criteria might be appropriate. While this approach may make use of the assessment framework easier, its overall effect is that the assessment is less intersubjectively verifiable. There are no documents available for either model that provide expert knowledge from previous applications, but some have published evaluation reports (Miller, Blake & Sorsby, 2012).

In terms of *prescriptive use*, AIDA does not provide specific *improvement measures* (3.1), no *decision calculus* for selecting improvement measures (3.2), and no *methodology for making these decisions* (3.3).

Table 8 summarizes the findings of the evaluation.

<table>
<thead>
<tr>
<th>Principle</th>
<th>Evaluation of AIDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Basic information</td>
<td>No <strong>prerequisites for applicability</strong> are noted. Higher education context is implied but not explicitly noted as a requirement. <strong>Class of entities under investigation</strong> is not defined. Elements address both department and institution levels and cover a range of entities such as processes, systems, and resources.</td>
</tr>
<tr>
<td>1.2 Concepts of Maturity</td>
<td>Maturity is not clearly defined but focuses on externalization and builds on the <em>Five Stages</em> and three-legged stool model (McGovern &amp; Kenney, 2003). Multiple <strong>levels of granularity of maturation</strong> are available through three legs, each with a number of elements. <strong>Theoretical foundations of change</strong> are left underspecified.</td>
</tr>
<tr>
<td>1.3 Domain-specific concepts</td>
<td>The illustrative nature of statements is reflected in informal descriptions of central <strong>domain constructs</strong>.</td>
</tr>
<tr>
<td>2.1 Intersubjectively verifiable criteria</td>
<td>Many criteria are subjective and illustrative, as acknowledged by the creators.</td>
</tr>
<tr>
<td>2.2 Assessment methodology</td>
<td>Basic procedure and some guidance provided.</td>
</tr>
</tbody>
</table>
In summary, the analysis of frameworks shows several characteristics indicating that key design principles were not taken into account in their design. To understand how the characteristics of the frameworks affect their use, however, we need to conduct an empirical examination of both models in a real-world environment. This is the focus of the next section, *Usage Evaluation*.

<table>
<thead>
<tr>
<th>Section</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Improvement measures</td>
<td>Not addressed.</td>
</tr>
<tr>
<td>3.2 Decision calculus</td>
<td>Not addressed.</td>
</tr>
<tr>
<td>3.3 Assessment methodology</td>
<td>Not addressed.</td>
</tr>
</tbody>
</table>
8 Usage Evaluation

To characterize frameworks in use, we use case study research to investigate assessment projects in real-world organizations. First we discuss additional details of the empirical method for evaluating the use of assessment frameworks. The study’s procedure and the data collected at each stage are described in detail with the intention that the same research method can be applied to study other frameworks. This is followed by four individual summaries of assessment projects (DPCMM in use at ASA, DPCMM in use at UTL, AIDA in use at ASA, and AIDA in use at UTL).

8.1 Case Study Design

Case study design requires a set of choices (Yin, 2014), and the design for this method is described accordingly.

1. What is the unit of analysis?

The overarching unit of analysis in our method is this assessment project, and the key interest is in the assessment framework (the set of assessment models, tools and guidelines) and its quality in use as the key component in that project. This means that the process, its results, and the framework constitute embedded units of analysis for the case study design.

2. Should the study focus on a single case or multiple cases?

While a single case study would allow some insights into the practice of assessment, multiple-case designs allow a comparison across multiple projects and organizational settings (Eisenhardt, 1989) and provide more robust analytical conclusions due to the ability to cross-reference and compare events to identify alternative explanations and increase external validity (Eisenhardt & Graebner, 2007; Yin, 2014). We have thus adopted a multiple-case design that enables a comparison of each framework across organizations as well as a comparison of the organizations’ contingencies across assessment projects.

3. Which sources of evidence should be used?

Our choice of evidence is motivated by the need to triangulate between direct observations, explicit statements made in the assessment process, and the reflective perception of the practitioners. The method thus uses a combination of documentary artifacts created through the assessment; video-recorded observations of the assessment; an entry questionnaire; and recordings of exit interviews after each assessment. All recordings are transcribed and analyzed.

4. Is the analysis explanatory, exploratory or descriptive?

The method aims, as a baseline, to provide a detailed description of the phenomena of interest to enable explanations of how and why specific events arise in the process of assessment, how these events influence the value of the assessment, and how the framework could evolve to increase its usefulness. Case study research benefits from the prior development of theoretical propositions (Yin, 2014), and this work correspondingly introduces basic concepts from research on organizational assessment and structured perspectives on the use of these frameworks in practice.

In the multi-case study design, multiple frameworks are analyzed across multiple organizations. In this instance, two organizations (ASA and UTL) provided the context and setting for evaluation. This resulted in four cases, with each combination of framework and organization accounting for one assessment project.

A pilot case was conducted to explore the adequacy of the units of analysis, and to test the protocol and data collection mechanisms. This case involved an assessment conducted with one of the organizations using the NDSA levels of preservation (Phillips, Bailey, Goethals, & Owens, 2013). Minor changes were introduced to the data collection mechanisms consequently.
Any application of this method will require a clear definition of roles in the beginning that considers the unique circumstances of application. A distinction must be made between the roles of practitioners and evaluators in the assessment projects (the object of study), and in the case study research. In our evaluation setting, the roles of practitioners and researchers are as follows. Two assessment projects were conducted with the Austrian State Archives (ASA), where author Hannes Kulovits was the head of the Digital Archive, and he acted as the Assessor. For the two assessment projects at the University of Toronto Libraries (UTL), Steve Marks is the Digital Preservation Librarian and acted as the Assessor. Both practitioners participated in the case study research by providing additional insights on their experience and reflections on the frameworks after the completion of the assessment projects. The remaining authors are academic researchers. During the assessment projects, their roles included: Assistant to the Assessor (notetaking and recording scores), and Observers (observing and making research notes). After each assessment project, some researchers conducted an interview. All academic researchers then reviewed and analyzed the recordings and documentation produced during the assessment projects. The practitioners reviewed and confirmed the case study analysis and findings, and validated our summaries.

8.2 Data Collection

One assessment project provides the scope of an individual case. Within this scope, the data collection activities of the case study method use a combination of instruments to provide evidence from multiple angles for analysis. Figure 5 illustrates the data collected per case.

For each case, a pre-assessment survey establishes a baseline understanding of the organizational context and perspective of the Assessor prior to conducting the assessment. The survey covers preservation concepts and principles, as well as the Assessor’s understanding of conceptual models, assessments or audits and is administered prior to the assessment project. Appendix 1 provides the questionnaire that is used.

During phases 2-4 of the assessment project, video recordings of the assessment process are created through web cameras and screen sharing and complemented with notes by the research team and the Assessor. The recordings show, at any point in time, all participants’ headshots, plus the screens of the computers used throughout the assessment for researching information and recording assessment results. All recordings are fully transcribed using a professional transcription service and retained for analysis. Through the assessment project itself, documentation generated by the assessment itself (scores, comments, results, etc.) is recorded and retained for analysis.

After the completion of each assessment, the research team conducted a semi-structured exit interview with the Assessor. This was done to capture the Assessor’s impressions of the framework, the results, and the overall process of assessment, and to provide room for reflection. Appendix 2 provides the set of initial questions used in these interviews. The interview is recorded and transcribed.
8.3 Analysis and Reporting

Analysis of each case was done using qualitative coding, which helped to identify specific aspects of interest in the observations, recordings and documents. The coding was done by two researchers who worked independently before comparing and iterating their results until consensus was reached. To facilitate analysis, the embedded units of analysis were used to structure accounts of each assessment project in a case report. The case reports and their findings are then validated by the Assessors.

Our study of DPCMM and AIDA includes four individual assessment projects, applying each of the two frameworks in two organizations. The four assessment case summaries are structured as follows:

- The assessment process is described and completed with a tabular summary of the effort and time involved and relevant events that occurred as the assessment progressed, as noted by the assessor or observed by the researchers.

- The assessment results and outputs are discussed, including the scores, areas of strength and weakness highlighted in the results, as well as the assessor’s impression of the value and usefulness as well as trust in the results.

- For each significant incident that was noted, a table summarizes when (process stage) and where (model component) it occurred, what happened, and which consequences it had.

The summaries are grouped by framework. A subsequent analysis section provides a summary of each framework’s quality in use across organizations. The detailed description of relevant events that occurred is provided as a basis for further analysis.

8.4 Case 1: DPCMM in Use at ASA

8.4.1 Assessment Process

Table 9 summarizes key characteristics of the assessment project. In many instances, the Assessor chose to evaluate a statement positively if the “system is capable of doing it” or it’s “theoretically possible” even though no active practices are currently in place, or the specific action or goal is not relevant to the ASA. Yet, on some occasions, e.g. Archival Storage (10), he chose to evaluate the current, not potential practice.

When components of the framework were focused on specific documents such as policies, procedures and strategies, the Assessor wondered how to interpret the degrees of formality that exist in his organization in light of the statements of the DPCMM. For example, he was unclear how the intentions of a draft document not yet formally approved should be assessed.

While reviewing his initial answers in Phase 3 (Review Assessment), the Assessor provided additional information to reinforce the selected statement and, for certain components, reconsidered and adjusted his selection of statements. His changes did not reflect re-considerations of the organization’s capabilities, but rather familiarity with the structure and format of the tool: Some questions involved variations in language like “most” and “all”, some involved “select all that apply” and others required the selection of a single statement as the best fit. This influenced the final results and scoring, which is important to note since this part of the assessment procedure was introduced by the research team and is not explicitly supported by the framework or tool.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Short summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time spent</td>
<td>The assessment was completed by the Assessor, with the aid of the note-taker. It progressed quickly, in two hours from start to finish. The Assessor</td>
</tr>
</tbody>
</table>
took approximately one hour to review the DPCMM documentation before beginning the assessment. The first pass through the all components was completed in approximately one hour, with a second pass for review completed in 40 minutes.

<table>
<thead>
<tr>
<th>Effort</th>
<th>5 person-hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resources used</td>
<td>Most of the time spent reading statements and explanations, DPCMM background documents and material in DigitalOK.</td>
</tr>
<tr>
<td>Skills and knowledge</td>
<td>Assessor possessed all knowledge and rarely struggled to recall information about the organisation. He frequently described a specific document or scenario to support his judgments.</td>
</tr>
<tr>
<td>Strengths highlighted in results</td>
<td>In Technical Expertise (5), Ingest (9), Device/Media Renewal (11), Integrity (12), Security (13), and Preservation Metadata (14), ASA achieved the highest possible rating. In Collaborative Engagement (4) and Access (15), the second highest score “Advanced” was achieved.</td>
</tr>
<tr>
<td>Weaknesses highlighted in results</td>
<td>In Policy (1), Strategy (2), Governance (3), and Electronic Records Survey (8), ASA achieved the level of Minimal, the second lowest score. In the criterion of Archival Storage (10), ASA was placed at the lowest possible rating of Nominal.</td>
</tr>
<tr>
<td>Difficulties</td>
<td>Difficulties in the use of the model surfaced at a number of components:</td>
</tr>
<tr>
<td></td>
<td>• Collaborative Engagement,</td>
</tr>
<tr>
<td></td>
<td>• Electronic Records Survey,</td>
</tr>
<tr>
<td></td>
<td>• Ingest,</td>
</tr>
<tr>
<td></td>
<td>• Archival Storage, and</td>
</tr>
<tr>
<td></td>
<td>• Access.</td>
</tr>
</tbody>
</table>

The Assessor saw the assessment process as being of particular value to an organization at the early stages of their development, for example when creating a tender document, a scenario he has experience with. He described the use of the results as a base line and referred to the possibility of conducting the assessment again at a later time. The quick process of assessment was also valued and the Assessor described it as “much easier to do than TRAC.”

### 8.4.2 Assessment Results

The DigitalOK online tool provided an Index Score of 38/60, which placed the digital preservation capabilities of ASA at the “Advanced” level. The Assessor found the results to be accurate as a reflection of his own knowledge and expectations. A major exception is the result in the Archival Storage (10) component, where he disagreed completely with the resulting lowest possible score and provided arguments for the highly advanced storage capabilities at the organization. He stated, “this one is definitely advanced, but in the results … the maturity level is one which cannot be true. So,… I would have had to tick more criteria but I didn’t know that.” The Assessor attributed the discrepancy to a lack of clarity in the statements, the absence of instruction on use of the tool and confusion over whether or not the criteria were meant to be cumulative. With this notable exception, the Assessor was generally quite confident in the results, based on the alignment of the results with the Assessor’s knowledge of the organization. He acknowledged that the results provided only one perspective based on this specific framework and that this was a limitation. He also noted that the evidence required for the assessment was not as rigorous as what would be required in a “real audit” for documentation. He expressed some hesitation in his ability to find adequate documentation to support what he had stated.
Nevertheless, he was quite confident in his assessment and knowledge, stating, “I know the evidence behind it.”

The Assessor saw the results as a generally valuable baseline and described DPCMM as “sufficient and helpful” in that the results showed areas for improvement. For example, he was able to recognize that the ‘technical side’ was a particular strength of ASA, while the “policies, strategies, documented processes” were not. The Assessor even discussed a plan to share the results of the assessment with his supervisor.

However, he stated that the model lacked guidance for understanding what the different levels mean, the differences between them, and the requirements to progress to the next level. This limited the usefulness of the assessment in relation to the Assessor’s initial intent, which was to use the assessment to help advocate for additional resources, particularly for the weaker organizational dimensions of ASA, as part of efforts to achieve certification as a Trusted Digital Repository.

Table 10 details key incidents that arose during the project.

<table>
<thead>
<tr>
<th>Framework location and description</th>
<th>Description of difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collaborative Engagement (4)</strong> addresses how the organization manages and fosters collaboration both internally and with a broader community. The criterion includes adherence to a “collaborative digital preservation framework.”</td>
<td>It was unclear to the Assessor how this “framework” is defined, if it is intended to be a formal document, and how that would apply. The Assessor described specific collaborations that take place with more formal conditions, such as collaborations with their IT Service Provider and Electronic Records Management System provider, but it was unclear what stakeholders are relevant to address.</td>
</tr>
<tr>
<td><strong>Electronic Records Survey (8)</strong> addresses the inventory of electronic records for “volume and scope of electronic records that will come into its custody.”</td>
<td>Certain aspects of this component seemed too prescriptive and not applicable for the work of the ASA. For example, the Assessor noted in their survey of electronic records storage that “volume is not a parameter for us nor location.”</td>
</tr>
<tr>
<td><strong>Ingest (9)</strong> addresses the extent to which an organization complies with ISO14721 (OAIS) by accepting SIPs and creating AIPs during ingest. The terminology used in the DPCMM tool switches from “Surrogate SIPs” to simply “SIPs”</td>
<td>There was a misalignment between how closely ASA follows the OAIS model and what this component expects. The Assessor distinguished “conforming SIPs,” understood to contain all content required of a SIP, from “compliant SIPs” as those that include all metadata required to comply with OAIS. He stated that SIPs at ASA are not fully compliant since they miss administrative metadata from records producers. The terminology switch in the model caused confusion. During the assessment, he was not concerned with the distinction between compliant and conforming when reading higher-level statements. The organization scores at the highest level 4.</td>
</tr>
<tr>
<td><strong>Archival Storage (10)</strong> addresses concerns such as the storage tiers used, the degree of automation in establishing a...</td>
<td>The Assessor had difficulty with the terminology and found it difficult to categorize the organization’s set-up with terms such as “near-line” and “off-line” for which no definition or examples were given. He also had problems...</td>
</tr>
</tbody>
</table>
Evaluating Assessment Frameworks in Digital Curation

chain of custody and operational statistics. For the latter, one requirement states that “Semi-automated capture of archival storage operational statistics are used to support systematic digital preservation planning.” with the structure of the component. The form of the first early statements begins with “Only one instance…” but shifts at higher levels. Aspects of this component seemed too prescriptive and not applicable to the work of the ASA. He struggled to choose the criteria with the best fit. He noted that the system is capable of capturing operational statistics but these are not currently used for systematic planning, and left these statements unchecked.

| Access (15) addresses access to Dissemination Information Packages (DIPs) by communities of “Users.” | The ASA does not provide direct access to external users, so the Assessor interpreted the statements using archivists as users of the system. He also interpreted the statements as true if their system was capable of providing certain DIPs, even if it is not the organization’s current practice to provide these (such as the provision of certain formats). He interpreted “Users” here differently from Designated Community (7), where he considered it synonymous with the OAIS term “Consumers.” |

8.5 Case 2: DPCMM in Use at UTL

8.5.1 Assessment Process

Salient characteristics of this assessment project are presented in Table 11. Applying DPCMM, which was developed for an archival context, to an academic library forced the Assessor to interpret concepts and terminology related to electronic records in more relevant terms.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Short summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time spent</td>
<td>Due to time constraints and availability, the assessment was completed in separate sessions on consecutive days. The first pass through the survey was completed in just over 2.5 hours on day one. The review of the answers and submission of the survey was completed in about 40 minutes on day two. Since the tool failed to generate the report for download, the research team contacted the creators of the model, who provided the final report by email, and a third session was needed to view and discuss the results. This session was completed in 30 minutes on day two.</td>
</tr>
<tr>
<td>Effort</td>
<td>7.5 person-hours</td>
</tr>
<tr>
<td>Resources used</td>
<td>The Assessor frequently referred to a printed copy of the documentation and the DigitalOK User Guide, made use of the tooltips in the online form to confirm definitions of terms, and consulted other documentation to confirm or clarify definitions. For example, he looked up online documentation from a TRAC audit he had previously completed at another related organization and searched for other standards referenced (e.g. the Department of Defense standard noted in the DigitalOK Background information survey).</td>
</tr>
<tr>
<td>Skills and knowledge</td>
<td>The Assessor generally was able to answer all questions with his own knowledge of the organization.</td>
</tr>
<tr>
<td>----------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Strengths highlighted in results</td>
<td>In the following criteria, the second highest level was achieved: Strategy (2); Designated Community (7); Archival Storage (10); Device/Media Renewal (11); Preservation Metadata (14); and Access (15).</td>
</tr>
<tr>
<td>Weaknesses highlighted in results</td>
<td>In the areas of Policy (1), Collaborative Engagement (4), and Integrity (12), UTL had the second lowest rating (Minimal). None of the components were scored with the lowest possible rating of Nominal.</td>
</tr>
</tbody>
</table>
| Difficulties | A few difficulties occurred during the process in the components:  
- Collaborative Engagement,  
- OS/TN Formats,  
- Archival Storage,  
- Governance, and  
- Electronic Records Survey. |

Most components followed a consistent structure where the Assessor selects all statements that apply to their organization. However, there were two components that required selecting the one statement that most accurately reflected the Assessor’s situation (Device/Media Renewal, 11 and Preservation Metadata, 14). The tool ensured that only one statement could be selected for Device/Media Renewal (11), but not for Preservation Metadata (14). The Assessor missed this variation and selected multiple statements for this component, from which the tool used the highest level for the score.

Overall, the online tool did not instill trust on the part of the Assessor. It did not support a wide range of navigation, returning to the “Background Information” screen was not possible, and he found it difficult to return to the first component of the assessment to begin the review of answers. The Assessor was concerned that some of his entries would be lost if he returned to earlier pages. This did happen for the final component, Access (15), which was not submitted so that all components could be reviewed.

### 8.5.2 Assessment Results

The DigitalOK online tool provided an Index Score of 37/60, which placed the digital preservation capabilities of UTL at the Advanced level.

The Assessor found the results somewhat inaccurate in that they did not capture the range of cases and varying capabilities that exist simultaneously within UTL, even within a single repository. He found that within the structure of the framework, there was no way to capture the variability and breadth of practices existing at UTL. He selected all statements that applied for any of the components where this variation exists. Yet, many of the components were scored or summed in such a way as to determine the overall level by taking the highest-scoring statement selected, not accounting for a range of varying practices. He stated, “without some instruction, and maybe there was instruction there that I missed, if you’re in this case where some of your activities are at a higher level and some are at a lower level, default to the low level or the high level. Without some instruction on how to handle that case, it’s a bit unclear.” This was particularly problematic in that the Assessor found the overall results were biased too high, especially for an assessor who wants to use the results to bring about change and improvements in the organization. He had little confidence in the results.

The results were not regarded as very useful for UTL or a similar organization. The Assessor found general value in the model and its scope of 15 components, but felt that statements within each component were too prescriptive and would have found more value in a model that emphasized the thinking and rationale behind each component, rather than the specific actions or criteria to meet. Part of the Assessor’s concern was with the scales used for the levels, at one point he described them as ‘flat’ and geared towards answers in the middle of the range. The scores that emerged were not severe enough to motivate changes or granular enough to provide a clear direction for improvement. The
specific results matched his expectations and understanding of his organization, further limiting their usefulness.

The Assessor indicated he would present results to others in the organization, but with disclaimers and qualifications for how to interpret them. In his view, this would be the case for any assessment results. In his view, the real value of such results lies in their role as a catalyst for discussion, conversation, and questions. This corresponds to his initial motivations for conducting an assessment at the time of creating a new Digital Preservation Policy Working Group at UTL. Generally, he would recommend it for an organization starting out in developing their repository, but considered that an organization in the very early stages of digital preservation would benefit from other conceptual models outside the realm of assessment, such as the Kenney and McGovern three-legged stool model and workshops. He noted that this process is especially useful for those starting out in digital preservation:

“I think there’s probably a lot of value there … to those people and organizations reading a list of fifteen words that are digital preservation related with examples under them, right, and saying: oh, I should think about this, I should think about governance, I should think about storage and security.”

However, in his case the results were not as valuable, as he stated:

“There’s always value to sit down and check all your premises … but it didn’t make me think about things any differently. … [or] wallop me with new revelations or anything.”

In short, while identifying the framework as useful for perhaps developing organizations or practitioners, the Assessor did not find it well-suited to UTL or himself as an individual.

Table 12 details key incidents noted during the project.

<table>
<thead>
<tr>
<th>Framework location and description</th>
<th>Description of difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collaborative Engagement (4)</strong> addresses the ways an organization manages and fosters collaboration both internally and with a broader community, particularly through a “collaborative digital preservation framework.”</td>
<td>The Assessor was dissatisfied with this component and discussed important collaborations within UTL that may not be measured, or relationships that might not be captured in a formal document because that may not be the appropriate approach to these relationships at this time. The Assessor suggested that the tool might be improved by providing a way to record such exceptions (“add an attribute”) for cases where the model doesn’t quite fit the organization, making it generally less prescriptive.</td>
</tr>
<tr>
<td><strong>OS/TN Formats (6)</strong> prescribes specific open formats for certain types of digital objects or records. The format component does not distinguish risk factors and only allows the Assessor to</td>
<td>The Assessor struggled with this component and noted that describing formats as “Open Standard / Technology Neutral (OS/TN)” is uncommon. He re-interpreted this to mean supporting use of a format where he was comfortable with “how much information is out there about it.” He took issue with the vague and general statements related to format and discussed that with video, where the use of particular codecs has more impact on preservation than a</td>
</tr>
</tbody>
</table>

---

3 Digital Preservation Management Workshops and Tutorial: http://www.dpworkshop.org
| **Archival Storage (10)** | addresses concerns such as storage tiers, the degree of automation in establishing a chain of custody, and operational statistics. The terms “near-line” and “off-line” storage tiers are not defined, and no examples are provided. | It was unclear to the Assessor what these might look like, or how they apply in UTL’s particular case, with their actual practices and systems that seems to cover some aspects or intentions of near-line and off-line, but also overlap. The Assessor interpreted some statements as how easy or quickly retrieval or recovery of information could be completed in the case of an emergency or loss. |
| **Governance (3)** | addresses the existence of and adherence to a “digital preservation governance framework.” | The Assessor identified underlying assumptions about the governance structure of the organization that are Archives-centric and expected to have specific mandates in regards to electronic records. In spite of this, UTL scored relatively high in this component and assessed at Level 3 “Advanced.” |
| **Electronic Records Survey (8)** | addresses ways to investigate, document and manage an organization’s electronic records, both for existing and projected “volume and scope of electronic records that will come into its custody.” The concept of a survey is framed around the existence of retention schedules, a document specific to archives. | He initially attempted to address the statements as-is but found it necessary to re-interpret the statements about “retention schedules” since UTL focuses on collections. He decided to interpret the statements through a mental “crosswalk” to “collections policies,” while maintaining what he saw as the question’s intention. The component is focused on “Electronic Records,” but the Assessor extended this to other digital materials being preserved. While making this judgement, he addressed the differences between retention schedules and collections policies and acknowledged that this is not an exact or perfect fit since each document would address risk differently. |
| **Digital Preservation Strategy (2), Collaborative Engagement (4), Designated Community (7), and Electronic Records Survey (8)** | contain quantifying statements including “all” conditions such as “The strategy calls for the transformation of all electronic records in native file formats to preferred preservation formats at ingest.” (Digital Preservation Strategy 2). | The organization’s heterogeneity meant the Assessor would not confirm statements addressing “all” conditions. The Assessor noted that the types of relationships with record producers and the diversity of records producers they work with make it impossible and, in fact, unreasonable, to have full 100% control or knowledge of their various record creation processes. Similarly, for questions around Designated Community and Access, UTL has structures in place to understand user groups, but could not cater to 100% of the needs of 100% of users as is demanded by the top-level statements found in these components. |
8.6 Case 3: AIDA in Use at ASA

8.6.1 Assessment Process

AIDA is tailored to higher education institutions, so applying the model to ASA required the Assessor to interpret how to apply university-specific concepts like funding or mandate to national archives. The ‘institutional level’ was applied to ASA as a whole and the ‘department level’ applied to the Digital Archives. The scores were reviewed after all elements in all three legs were completed. This resulted in changes that increased the scores for three components, as described further in the events noted below.

In many of the technology elements, the Assessor struggled with the descriptions of Stage 5 and his own expectations that the ASA had achieved a high level of maturity in technology capabilities. In Technology (4), he began by looking only at the higher levels saying “I would expect us to be in stage five here, so I’m reading four and five.” However, the kind of externalization specified - reporting to external partners or consortia members - did not apply to ASA. He similarly struggled with the Stage 5 definitions in Technology Elements Metadata Creation (10) and Institutional Repository (11) that prescribe external partnerships and membership in larger consortia.

The Assessor noted that considering some organizational and resources aspects in depth through this process was valuable. In comparison to DPCMM, a greater number of elements had to be considered, as well as the distinction between levels for the institution and department. The Assessor felt that the two-level split between institution and department made AIDA “more complex” and made it harder “to decide which stage to pick.”

Table 13 summarizes key characteristics of the assessment project.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Short summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time spent</td>
<td>The assessment was conducted by two people (the Assessor and a note-taker) in a single day. The full duration of the assessment was about four hours including breaks. A few minor pauses were required due to technical difficulties and connectivity issues since note-taking documents were on a cloud-based platform.</td>
</tr>
<tr>
<td>Effort</td>
<td>7 person-hours</td>
</tr>
<tr>
<td>Resources used</td>
<td>Most of the Assessor’s time was spent reading the criteria and determining how to apply or compare each to the situation at ASA. The Assessor used a paper copy of the elements for reference, and the note-taker recorded the final stage (1-5) as it was determined for department and institution for each element. Notes on the rationale for these scores were also recorded by the note-taker and confirmed by the Assessor before moving on to the next element. The worksheets and templates were simple word documents and easy to use, with the exception of the weighted scorecard. Provided as a spreadsheet that calculates the overall score at the end of the assessment, it was initially unclear to the Assessor and note-taker how to fill in the scores and how the sums should be interpreted. No supporting documentation is provided for the scorecard.</td>
</tr>
<tr>
<td>Skills and knowledge</td>
<td>The Assessor’s existing knowledge and expertise contributed to the short duration of the assessment since he was able to recall most information</td>
</tr>
</tbody>
</table>
required for each element from memory. The only exceptions were the elements covering financial auditing and policy review. In these cases, the Assessor was aware of activity in the area, but had limited knowledge of details. This gap in expertise was discussed during the exit interview and it was noted that consulting with other staff would be necessary to provide more detailed and nuanced responses.

### Strengths highlighted in results

The assessment revealed that ASA’s highest ratings were in the Technology leg. Both the department and the institution were Stage 5 in: Ensuring Availability and Integrity (3), Integrity of Information (4), Obsolescence (5), Security of Environment (7), Security Mechanisms (8), Implementation of Disaster Recovery Plan (9), and Institutional Repository (11).

### Weaknesses highlighted in results

The Resources leg was the weakest area for ASA, notably in Staff Development (9) and Review of Business Plan (2), where both the department and the institution rated poorly.

### Difficulties

A few difficulties occurred during the process in the components

- Organization (4) “Asset management and sharing”,
- Organization (6) “Audit trail of activities and use”,
- Technology (1) “Technological Infrastructure”,
- Technology (2) “Appropriate Technologies”,
- Resources (5) “Transparency and auditability”, and throughout Technology on level 5.

#### 8.6.2 Assessment Results

The department scored equal to or higher than the institution in nearly all elements. Differences were identified in Risk Analysis (4), where the institution had the lowest rating, while the department had the second highest, and Staff Numbers (8) which rated the institution higher than the department.

The Assessor found the results were generally useful for their coverage of a range of areas addressed by the model and the depth with which it examined them. In particular, AIDA served to call attention to the need for further development related to criteria in the Organization leg, which matched the Assessor’s intuition about ASA’s strengths and weaknesses. However, he felt insufficient guidance was provided on the interpretation of the results, which limited the value of the report.

He expressed modest confidence in the results of the assessment, based on the relative agreement between the results and his sense of ASA’s strengths and weaknesses. He perceived the results to be fairly accurate and expressed a willingness to share the results with colleagues and supervisors within the organization, provided they were presented with sufficient context. However, he would not share the results publicly. He noted that for some elements in Resources, additional input from other staff would help improve accuracy. In “Asset Management and sharing” (Organization 4), the concept of sharing, re-purposing, and re-use was unclear to the Assessor, and therefore he was not confident in the score.

Table 14 summarizes relevant incidents noted during the project.

<table>
<thead>
<tr>
<th>Location and description</th>
<th>Description of difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization (4) “Asset management and sharing”</td>
<td>The Assessor had difficulty interpreting this component in the ASA’s context. The statements for Stage 4 note that</td>
</tr>
</tbody>
</table>
Evaluating Assessment Frameworks in Digital Curation

<table>
<thead>
<tr>
<th>Organization (6) “Audit trail of activities and use”</th>
<th>“Re-use and repurposing” is either possible or widespread, which is not part of the ASA mandate as an archive. During review, the Assessor reconsidered his initial interpretation, deciding that the use of unique identifiers facilitates re-use and sharing. The score was changed from level 3 to 4 for both institution and department.</th>
</tr>
</thead>
<tbody>
<tr>
<td>In assessing the institution, the Assessor found the statement for Stage 4 confusing. It requires that “the history can be used to generate reports to meet targets or comply with certain external requirements.” He expressed confusion over whether the element is determining what is audited in practice, or what potentially could be audited by the system if/when the need arises in the future.</td>
<td></td>
</tr>
<tr>
<td>Technology (1) “Technological Infrastructure”</td>
<td>When looking at the institution, the Assessor did not see the distinction or clear progression between Stage 4 and 5. Stage 4 states: “Infrastructure is sufficient for the organization’s needs…Upgrades and enhancements are planned and budgeted for.” Stage 5 focuses on externalization and describes institutional infrastructure as “highly integrated” and linked to “external services.” He also noted that it was difficult to translate the meaning of this kind of externalization to ASA’s situation. It was generally unclear to the Assessor how one progresses from Stage 4 to 5, therefore the Assessor chose Stage 4.</td>
</tr>
<tr>
<td>During the review, the Assessor changed the score from 4 to 5 for the Department, deciding that “on the department level, this could be seen as externalized because tenants are outside department and outside the organization.”</td>
<td></td>
</tr>
<tr>
<td>Technology (2) “Appropriate Technologies”</td>
<td>The Assessor had limited knowledge of the details of financial practices at the institutional level. While he was certain of the centralization and existence of auditing practices, the details of audits would have to be verified with a colleague from that department. Originally, the Assessor read the more general statements for each stage and noted that the concept of a “local audit” fit for the ASA. This was re-evaluated during review from Stage 2 to 4. Though the Assessor could not confirm details of auditing, he was sure practices were centralized and transparent.</td>
</tr>
<tr>
<td>Resources (5) “Transparency and auditability”</td>
<td>---</td>
</tr>
</tbody>
</table>
8.7 Case 4: AIDA in Use at UTL

8.7.1 Assessment Process

Table 15 summarizes key characteristics of the assessment project, and Table 16 details the difficulties that arose. For the two levels AIDA assesses, the ‘institutional level’ applied to UTL and the ‘department level’ applied to the IT Services unit. This distinction seemed to fit well and make sense.

For each element, the Assessor considered the different indicators and exemplars given for the institution and department. These were generally found to be useful in the assessment process to locate the organization at a specific Stage. Using these, the Assessor would quickly decide if they were above or below a given stage. In cases where the Assessor was debating between two stages, he chose the lower stage as noted in the instructions for AIDA.

No changes were made to the scores during review. The Assessor noticed overlap for disaster planning between the different legs and marked these elements for detailed review. The distinction in what was being assessed in the separate elements for Organization (11) “Disaster planning and business continuity” and Technology (9) “Implementation of disaster recovery plan” was only apparent to the Assessor after going through the whole assessment. After reviewing, he decided to leave the initial score and interpretation, which focused more on business continuity than disaster planning.

The Assessor found that undertaking the assessment was a worthwhile exercise. He remarked that AIDA “struck a good balance between being granular enough… without making it overwhelming” and emphasized the model’s flexibility and provision of exemplar statements. He liked that the framework emphasized the expertise and responsibility of the Assessor in knowing the organization and encouraged informed interpretation of the criteria “rather than this sort of… ‘judgment will be passed on you based on the answers you give.’”

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Short summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time spent</td>
<td>The assessment was conducted over two days and took under three hours to complete. The Organization leg was completed on the first day, with the remaining two legs and the review process conducted on the second day. Some short breaks were required due to technical difficulties unrelated to the assessment tools.</td>
</tr>
<tr>
<td>Effort</td>
<td>7 person-hours</td>
</tr>
<tr>
<td>Resources used</td>
<td>The Assessor used a paper copy of the elements for reference and the note-taker from the research team filled in the electronic templates to note the stages (1-5) for both institution and department for each element. The rationale for these scores were also noted and confirmed by the Assessor before moving on to the next element. The worksheets and templates were simple word documents and the Assessor found them easy to use. He used a printed copy of the elements and made notes directly on the table for each element.</td>
</tr>
<tr>
<td>Skills and knowledge</td>
<td>The Assessor recalled all relevant information required to complete the assessment. More details and examples were provided for the department level, since this was the area of focus for the assessment and the practitioner.</td>
</tr>
<tr>
<td>Strengths highlighted in results</td>
<td>In the Resources leg, the highest stage was achieved by UTL for the following elements: Business Planning Process (1), Review of Business Plan (2), Transparency and Auditability (5), Sustainability of funding (6), Staff</td>
</tr>
</tbody>
</table>
38 | Evaluating Assessment Frameworks in Digital Curation

<table>
<thead>
<tr>
<th>Weaknesses highlighted in results</th>
<th>The institution and the department both scored poorly on the Audit Trail of Activities and Use (Organization 6), the Obsolescence and Changes to Critical Processes (Technology 6), and Risk Analysis (Resources 4).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulties</td>
<td>Difficulties arose in</td>
</tr>
<tr>
<td></td>
<td>• Organization (1) “Mission Statement”,</td>
</tr>
<tr>
<td></td>
<td>• Technology (5) “Obsolescence”</td>
</tr>
<tr>
<td></td>
<td>• Resources (1) “Business planning process”,</td>
</tr>
<tr>
<td></td>
<td>• Resources (8) “Staff Numbers”,</td>
</tr>
<tr>
<td></td>
<td>• Organization (6) “Audit trail of activities and use”,</td>
</tr>
<tr>
<td></td>
<td>• Organization (11) “Disaster recovery and business continuity”, and</td>
</tr>
<tr>
<td></td>
<td>• Technology (9) “Implementation of disaster plan.</td>
</tr>
</tbody>
</table>

8.7.2 Assessment Results

The weighted scorecard results revealed the institution scored higher than the department in the Resources leg, while the department scored higher in the Technology leg.

Overall, the assessment was seen as valuable. The Assessor remarked that he would be “confident saying to people in my organization: hey, this was a useful exercise.” However, the results did not identify discrete targets or objectives for improvement, but only provided broad indications of the strengths and weaknesses of different areas. The Assessor noted the difference in scores for the institution and the department for specific elements or legs, and generally agreed when one was at a higher stage than the other.

However, the presentation of the results, and the format of the weighted scorecard were seen by the Assessor as less significant than the exercise of undergoing the process of assessment. In order to look closer at the relative weighting, the Assessor examined the formulas in the spreadsheet itself, and then concluded that this made sense, with the “big picture” pieces weighted more strongly. Still, he noted that understanding the relative scores between department and institution was more useful, particularly to begin discussions with other UTL stakeholders. The weighted scorecard provides an overall score out of 5 for each leg, but the Assessor did not find it very useful, remarking “I’m struggling to find any meaning in 4.4 versus 4.2.”

The results of the framework were generally in line with the Assessor’s expectations and seen as an accurate representation of the organization. He noted, “this feels pretty solid to me actually… this rings true to me.” The relative scores across and within the legs of technology, resources and organization seemed appropriate to the Assessor. He had modest confidence in the results, recognizing the potential for inaccuracy in his own scoring.

Table 16 lists key incidents noted in the project.
<table>
<thead>
<tr>
<th>Location and description</th>
<th>Description of difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organization</strong> (1) “Mission Statement” assesses the degree to which the organization has a formal mission statement that addresses the management and creation of digital assets.</td>
<td>The Assessor was not sure if a Mission Statement document exists for the University as a whole. The Assessor found it more useful to consider the UTL strategic plan since it includes statements addressing commitment to curation of digital assets. It was unclear if this was an appropriate substitution since the model does not specify what needs to be addressed by a “mission statement.”</td>
</tr>
<tr>
<td><strong>Technology</strong> (5) “Obsolescence” addresses issues of obsolescence, referring to obsolescent materials, media, and file formats.</td>
<td>The Assessor attempted to find explanatory notes on this element but none are provided. He stated his own interpretation and definition of the element’s intention as the “technical ability to identify and deal with formats and media types that are obsolete.”</td>
</tr>
<tr>
<td><strong>Resources</strong> (1) “Business planning process” addresses concerns of sustainability of assets, particularly addressing funding and revenue.</td>
<td>The criteria in the higher stages place significance on income generation, but due to its role within the university and organizational objectives, UTL is rarely engaged in income generation.</td>
</tr>
<tr>
<td><strong>Resources</strong> (8) “Staff Numbers” addresses the availability of staff to support asset management.</td>
<td>The Assessor found it difficult to distinguish between stages 3 and 4 and interpret the difference between the phrases “Department has appropriate and sufficient staff with sustainable funding” (Stage 3) and “Funding enables the Department to keep staff numbers at a suitable level” (Stage 4). He noted that this distinction makes more sense for an organization whose mandate is not specifically about asset management. For UTL, this distinction is less clear, since, as the Assessor stated, “good asset management is a core part of our operations, not something that we do with our excess capacity.”</td>
</tr>
</tbody>
</table>
| **Organization** (6) “Audit trail of activities and use” addresses the transparency and accountability of actions and activities involved | The Assessor initially interpreted this element in terms of financial auditing, but reading the examples he realized that the focus is on the ability to track asset creation and use. He noted, “I recognize
<table>
<thead>
<tr>
<th>in asset management.</th>
<th>and am appreciating that the information provided here is not prescriptive, just illustrative in nature, but they’re useful illustrations.” The Assessor then repeated his initial reactions about financial auditing (that there are compliance requirements for a public institution) when assessing in Resources (5) “Transparency and auditability.”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disaster Planning:</strong>&lt;br&gt;<strong>Organization (11)</strong> “Disaster recovery and business continuity” addresses availability of an institution-wide disaster plan, <strong>Technology (9)</strong> “Implementation of disaster plan” addresses implementation and testing of business- and service-continuity.</td>
<td>The Assessor expressed minor confusion distinguishing these two related concepts. The Assessor’s initial understanding of the Organization element was reconsidered later on in the process, when he reached the related element in the Technology leg. However, no changes were made to the scoring, despite the new interpretation of the element.</td>
</tr>
</tbody>
</table>
9 Analysis of Assessment Projects

The questions that motivated the empirical evaluation of assessment projects focused on (1) resources and skills, (2) value and use for the organization, (3) accuracy and reliability of the assessments, (4) difficulties in the process and their resolution, and (5) the characteristics of the assessment framework as they surface in the assessment project and its outcomes. We will structure our discussion of findings accordingly.

9.1 Resources and Skills

All four assessments were completed in a few hours by a team of two people - the Assessor with responsibility and deep knowledge of the organization, and a note-taker well-versed in digital preservation terminology and generally familiar with the organization. Additional time (approx. 1 hour each) was required for the Assessor to review the framework’s documentation and become familiar with the framework prior to beginning the assessment. The low overall effort suggests that in organizations with comparable levels of well-established capabilities, a knowledgeable assessor and note-taker, the assessments can be completed in a single day with limited use of external resources such as access to colleagues in different departments across the organization or the documents describing their activities. The Assessor’s expertise in the discipline and familiarity with the conceptual foundations of the models meant that little time and effort was spent learning about these as parts of the assessment. The presence of a note-taker familiar with technical terms and abbreviations and the organization’s context allowed the assessor to focus on the assessment, with virtually no need for internal clarification of terminology. This was an efficient and recommended arrangement. Since the Assessors at both organizations are responsible for building and managing the assessed capabilities in their respective organizations, they saw virtually no need to verify their knowledge through consulting evidence such as process documentation, observation, or interviews, as is common for audits (Ross & McHugh, 2006). One participant remarked, “I don’t need to present evidence to myself.”

9.2 Value and Usefulness

The process of assessment itself was found to be useful and interesting to the Assessors in each case. However, this perceived value owes more to the general activities of reflection than the models. When the Assessors reviewed the reports generated at the end of the assessment process, they found few new insights or surprises. The reports largely reflected their own understanding of the organization, and little further detail was provided beyond comparing scores or weights. Both models revealed strengths in the technical aspects of both organizations, and similar areas of weakness, specifically related to policy, business plans, changes to critical processes, audits and governance. Where these models differed, it was in aspects of digital preservation that were covered in finer granularity with AIDA and not explicitly addressed by DPCMM. For example, AIDA found weaknesses in risk analysis and certain aspects of staffing, and strengths in disaster recovery at both organizations. These are not considered in the DPCMM model.

None of the models provided valuable insights for concrete steps towards improvement. For DPCMM, one Assessor noted the resulting report “doesn’t really give me a lot to work with outside of what I gave it.” AIDA provided results that allow for greater comparison between the different legs as well as departmental and institutional levels. However, neither framework provided much guidance for improvement and did little to specify steps to take towards achieving the next stage of maturity: “[T]he input that I’m getting back is you need to check more of the boxes.” In contrast to the resulting scores, the process was valuable since it required the Assessor to systematically think about multiple aspects of their organization’s capabilities. Even when the Assessor felt confused by, or disagreed with the concepts defined in the framework, the exercise itself was considered useful as a trigger to reflect on their own knowledge and assumptions.
9.3 Accuracy and reliability

Since neither assessment framework required documentation or evidence to support the assessor’s statements and judgments, the conclusions we can make about accuracy - the correct representation of an organization’s real abilities in terms of the model’s score - are limited. Without conducting a full audit, accuracy can only be considered in relation to the assessor’s perception of how consistently the framework reflected their own expertise and intuition, and how far it allowed them to express their judgment accurately.

While accuracy thus cannot be fully determined, it is clear that none of the assessment projects were found to provide particularly reliable scores. Several incidents in the assessment process pointed to a lack of reliability in that different assessors, or even the same assessor at a different point in time, may interpret the same criteria set differently due to a lack of clarity in its terminology and, sometimes, a lack of clear and distinct thresholds. This was evident in changes made during the review stage, when the same assessor interpreted the criteria in different ways, without noticing the shift in interpretation.

Without independent verification of evidence, as is done in a full-scale third-party audit, we can and should not expect high degrees of reliability. However, it was noticeable that inconsistencies in structure and terminology evident in the design of the models gave more room than necessary for subjective interpretation in different criteria and statements. This corresponds to reports that claimed AIDA is “possibly leaving too much room to interpret what is supposed to be evaluated” (Miller et al., 2012, p. 96).

9.4 Difficulties and opportunities

The assessment projects were helpful in identifying potential areas of improvement for the frameworks by highlighting difficulties in the assessment process. Some arose in scenarios that could be considered ‘typical’, in domains for which a framework was designed, others in ‘untypical’ contexts of use where a framework was applied in a somewhat different domain or environment. The difficulties that arose throughout the four assessments were brought up by the Assessors, but also independently observed during the assessment process.
10 Analysis of Assessment Frameworks

The recurring events in these projects highlighted characteristics of the frameworks and several themes emerged. The following sections will discuss each framework and how these themes arose with the framework in use. We structure the discussion according to questions of terminology, structure, the concept of maturity, gaps that emerged in the process, and the suitability of the framework and its flexibility across different contexts.

10.1 Analysis of the DPCMM in Use

Generally, the framework’s approach was found to be very prescriptive of specific scenarios, predicated on assumptions about organizational contexts and emphasized through the checklist approach of the online tool. While this approach can have benefits, the inconsistency in terminology and logical structure made it challenging to apply DPCMM effectively.

**Terminology:** The assumed or implied rigidity of a structured checklist of statements in DPCMM is contradicted by the ambiguity of the language used in each statement. Some terminology was adopted from domain-specific standards like OAIS, but not consistently. For example, the term “Users” is applied inconsistently in two components. Designated Community (7) employs the term “Users” instead of the standard OAIS terms “Consumers.” This is confusing since the other of the two important terms (“Producers”) is used consistently with the OAIS definition. In Access (15), the meaning of “access to DIP” is not fully defined, but implies external access and equates “Users” (Consumers) with system users. The glossary includes an entry for “Producers” that references the OAIS correctly, but not for “Users.” It includes “Consumer,” but the term is not used. This inconsistency in drawing from the OAIS Reference Model was revealed in use when the term “Users” was interpreted both as synonymous with the OAIS-defined “Consumers” and as a different type of system user internal to the organization. The inconsistency in interpretation could presumably be rectified through more consistent terminology, a more rigorous glossary, or better examples.

Similarly, definition of general terms like “most” were unclear. At UTL, the Assessor stated his assumption that this would mean greater than half. The open-ended nature of these terms may be an intentional feature allowing the Assessor to define what is acceptable or appropriate to their situation, but there is no room or guidance to document this. A clear definition would allow for more reliable judgements across different assessments.

**Logical Structure:** A prominent difficulty for both Assessors was in understanding and negotiating the inconsistencies in the logical structure of the model, which was often reflected in the tool.

At ASA, the Assessor noted that despite the overall ease of use of the tool and model, the selection of statements was inconsistent across components and caused confusion as to whether statements were cumulative (select all that apply) or separate (choose one): “sometimes you only have to tick one of the criteria and sometimes you have to tick more - this is not obvious.”. The differences between these types were not supported by a distinctive interface. At UTL, the Assessor also noted the inconsistency between choosing one statement that fits best, or all that apply. For example, the tool’s interface for component Device/Media Renewal (11) is the only component that constrains the user to one selection. Statements that captured specific quantities were especially confusing with the “select all that apply” structure. For example, in “Open Source / Technology Neutral Format (6),” if an organization has adopted three formats, it logically follows that they must have adopted at least one format. Yet, it was unclear how the tool would calculate the score for this component if the Assessor had not selected the statement for the lower quantity.

This was further complicated by the inconsistency in the language of the statements. At UTL the Assessor noted that the statements and levels are “changing multiple variables at once,” pointing to a lack of granularity. This was noted for Integrity (12), where each statement changes both the type of hash used for integrity checks as well as the timing for generation of the hashes. Separate statements...
for each of these concerns could provide the same results for determining the overall level, but would be easier for the Assessor to understand and consider in terms of their own organization.

The Assessors (at UTL especially) seemed to be seeking reasons for these inconsistencies in the model, expecting that when a component did not follow the common structure there was something exceptional about that component in particular.

**Maturity:** The concept of maturity and the definition of the levels of maturity in the DPCMM were problematic for both cases. At ASA, the achievements described at the “optimal” level were insufficient from the perspective of the Assessor, especially in the technical components (9-15) since these must account for constant changes to the technical environment. He proposed incorporating concepts of further optimization or continuous review to the Stage 5 requirements. He noted:

> ‘The highest maturity level in my opinion should be something that is working on further optimization, you know that means okay take a look at your processes and how good are they, maybe can you improve something, or are you doing too much … in some of the areas we have the highest maturity level but I know we have these processes now but we are not really reviewing them continuously. So this should … be reflected in the maturity level.’

At UTL, the highest level statements were not considered adequate for the highest goals the organization could achieve, and aims to achieve. The Assessor noted that there should be more room to the top: “there are certain pieces of this - digital preservation strategy being one that jumps out at me - I feel like we are not where I want to be, so I guess I’m a little surprised in cases like that could be rated so highly.”

**Gaps:** Both Assessors noted gaps in how the model addresses business processes and governance. At ASA, the Assessor found that the model did not address the implementation of business processes and instead focused on key documents such as policies or frameworks. The Assessor found the model provided no view of “how the organization has really implemented its processes.” In this sense, the overall capabilities and maturity of an organization are not only about achieving specific goals or conforming to a specific reference model, but also ensuring that processes are more widely understood and managed. At UTL, the Assessor noted in the exit interview that the model’s coverage of issues related to Policy, Strategy, and Governance overlooked important questions of mandate, and that motivations that impact succession planning and organizational sustainability are missing:

> ‘one thing that I worry a lot about in my work is sustainability…. there’s sustainability in terms of technology and making sure that your technological foundations don’t rot away by refreshment and review process and things like that, but there’s no real question about organizational sustainability … the closest we get is talking about mandate … formal strategies, and governance. One of the things that I tend to think is really important is … is this something that we’re doing because we have the mandate to do it and that means we have dedicated resources and … either a regulatory or a legislative reason we’re allowed to do it or is it just something we’re doing out of the goodness of our heart, and if it’s the latter case that raises a lot more questions about how to deal with things like succession planning and stuff like that, which … wasn’t mentioned here. So, I get that those things may be outside of scope, but they’re not, right?’ (UTL Assessor, 2015)

**Suitability and Flexibility:** The DPCMM claims to be applicable not just to archives, but also libraries and museums. While the DPCMM is seen as generally suitable for the ASA, the deployment at UTL shows the limitations of its flexibility. Unfortunately, the model was not flexible enough to capture the nature of the organization and its breadth of users, records producers, and collections. The variability in practices described by the Assessor could not be captured within the model. The organization’s heterogeneity also meant the Assessor would not confirm summative statements mandating a uniform approach to “all” instances of a certain criterion. In short, the Assessor did not find DPCMM to be suitable for use at UTL or flexible enough for thorough application.
10.2 Analysis of AIDA in Use

The approach of the AIDA framework encourages interpretation of criteria and thus relies more on subjective judgments and active interpretation of criteria by the Assessor. This approach is emphasized in the way the toolkit describes the different stages for each element, many using exemplars or illustrative statements rather than prescribing specific requirements that must be met. The overall goals of the toolkit are to facilitate communication between different groups rather than provide a reproducible outcome. This was appreciated by the Assessors and should be considered in the evaluation.

As with DPCMM, we structure our analysis of the framework into questions of terminology, structure, maturity, evident gaps, and suitability and flexibility.

Terminology: While some terms may be subjective and interpreted for a specific organization, the absence of a glossary was noted by the Assessor at ASA. For Organization elements (2 and 3) concerning policies and procedures, the Assessor found it difficult to tell the differences between defined procedures, a “policy” or a “real policy document.” For example, in the ASA’s organizational context, there are different stages a policy document might go through: release in draft form; under review; and officially issued. At UTL, the Assessor noted that for Organization element “Mission statement” (1), the qualities or requirements of a mission statement document are not defined. This made it difficult to determine if another type of document could be appropriately substituted to meet the intent of the element.

Logical Structure: While a linear process was not strictly dictated (as opposed to the structure of the tool for DPCMM), the structure of the model generally directed the assessment to proceed through the elements in each leg in order. While the toolkit specifies connections between different criteria across legs, there is little opportunity or ability to cross-reference elements during the assessment itself. At ASA, AIDA gave the Assessor the impression of logical inconsistency in the progression of some of its stages. One element in particular, Resources (1) “Business planning process”, did not appear to be incrementally progressive to the Assessor. Unlike the stages in other elements, the indicators at the lower stages were not cumulatively included at the higher stages since, in ASA’s case, sustainable funding is available without a formal “business plan” document.

Maturity: AIDA’s maturity levels are based on the Kenney & McGovern Five Organizational Stages of Digital Preservation (2003): (1) Acknowledge, (2) Act, (3) Consolidate, (4) Institutionalize, (5) Externalize. This places emphasis on actions or processes that extend beyond institutional boundaries and was problematic for both cases.

At ASA, externalization did not fit with the organization’s mandate in many ways. This was particularly problematic in the Technology leg, where the Assessor felt that the institution was at a very high level of maturity, but the exemplars for Stage 5 did not apply or make sense since they referred to external partners and consortial agreements. The model does not specify Stage 5 indicators for the Department for many elements, but for the ASA it makes sense that the Department (the Digital Archives) might ‘externalize’ by connecting to the broader archives and digital preservation community. Despite this omission, the Assessor scored ASA at Stage 5 at the department on a number of occasions and provided explanations of how the department had externalized its operations.

At UTL, the goal of externalization was often a better fit, but the Assessor noted the inconsistency in what the highest level (Stage 5) was measuring: “so we’re externalizing, in some cases that means we're part of a community and we're using established standards, in some cases it means that we're sharing data or metadata, in other cases it means that we’re being audited or that we are sending people out to give classes rather than just taking them. Like those all mean different things.”

Gaps: The only gap noted was the absence of a Stage 5 indicator or exemplar for many elements. This was most often the case for Department, but Stage 5 was also missing for the Institution for one element (Technology 6 “Changes to Critical Processes”). This was particularly prominent when assessing ASA.

Suitability and Flexibility: The AIDA toolkit was developed for a very specific context of use in higher education institutions in the UK. While UTL as an academic research library falls within the
higher education domain, focusing the assessment on UTL also required some adjustments and re-interpretation of the elements since UTL is not an academic department. Since preservation and asset management are a part of UTL’s core mandate, many of the exemplars were not applicable. The distinction between Institution and Department is based on the higher education context of the UK, but it was mapped easily to both organizations and was easy to interpret for most elements. Two elements, Resources (1) “Business planning process” and (2) “Review of business plan,” were a poor fit at both organizations. At UTL, these two levels did not always fit with the deep hierarchy and complex governance structure for an institution the size of the University of Toronto, and additional levels (to represent the overall university, UTL, as well as ITS) would have been beneficial for elements addressing funding and business plans (Resources, 1 and 2, concerning business plans).

At ASA, the usage of AIDA outside of its intended domain raised the issue of the limitations of the supporting documentation, which impacted its flexibility in use. Since ASA’s mandate is based on legal requirements, income generation was not a concern. Elements addressing the aspects of sharing, re-use and re-purposing were not applicable for ASA (Organization, 4 “Asset management and sharing”). Intellectual property (Organization, 10 “Intellectual Property Rights and rights management”) was also not considered relevant for ASA since the records are owned by the Republic of Austria. ASA’s concerns around access and rights are more focused on privacy and redaction of sensitive information. In general, these issues did not impact completing the assessment. However, a clearer articulation of terms and applicability would have better supported adaptation of the model across organizational settings.
11 Synthesis of Evaluation

We analyzed the empirical findings of the application of these models in their real environment. We focused on issues with the quality of the models that were uncovered in their use and show how design principles can explain these events.

11.1 Explaining incidents using design evaluation results

The assessors encountered a number of difficulties in their assessment projects and these drew attention to particular characteristics of the frameworks. While many of the characteristics that led to these difficulties could have been discovered through a close reading of the supporting documentation, limitations from the perspective of an assessor are not often noticed. The analysis described in the design evaluation sections above can support the identification of potentially problematic characteristics through its structured set of principles, but is unlikely to be comprehensive. It is in the assessment process itself that the consequences of product characteristics emerge. This aspect of the study links the analysis of the frameworks as artifacts with the empirical analysis of their use in real environments.

In the analysis of real cases, generally multiple events and consequences are related to a set of design principles. As part of the comparative case study analysis, we analyzes sets of related events across the four cases to explain how quality in use was influenced by model design. We progressed iteratively and cycled between an in-depth analysis of the recorded observations, statements, and critical reflections of the practitioners conducting the assessment, and the theoretical basis provided by design principles and methods. In this process, we carefully considered rival explanations such as the unique contingencies of the case study context and the expertise of the assessor. We focused on concrete instances where the characteristics of the assessment framework played a key role in influencing the practice of assessment. The discussion of selected events below focuses on those that provide particularly insightful observations and lessons learned, but it should not be taken as a comprehensive evaluation or validation of the models. Such an evaluation, however, should be conducted as part of a design process in which successive iterations of a model are built and evaluated (J. Becker et al., 2009; de Bruin et al., 2005; Peffers et al., 2007).

In discussing selected sets below, we will trace the characteristics identified through the models’ use back to their origins. We begin by explaining the events and their effects that revealed the framework’s quality in use, and connect these back to characteristics revealed in the analytical evaluation of design principles and the design process.

11.2 Interpreting and applying criteria and terminology

In both cases that applied the DPCMM, the assessors encountered difficulties with component 8 of the DPCMM, Electronic Records Survey.

At UTL, the assessor found it necessary to re-interpret the specific statements about “retention schedules” through an ad-hoc ‘crosswalk’ to a more library-centric concept of “collections policies”. He acknowledged that this is not an exact or perfect fit, as each document (i.e. collection policy vs. retention schedule) would approach risk differently. As a result, the assessor spent some time and effort to conceive of this re-interpretation, and acknowledged it was still not a great fit.

At ASA, the assessor had difficulty in evaluating certain statements in this component, determining they were not applicable: “volume is not a parameter for us nor location.” As a result, he struggled to respond since the criterion was not relevant and not applied. In the end he chose not to select the statement for this criterion. The resulting score was ‘minimal’, and it was unclear how this was influenced by a criterion with no relevance.

The specific characteristics of the product highlighted in these events show how design principles such as clear terminology and advice on adaptation can help to identify design aspects that could be
improved. The effects of these events directly impacted the efficiency, reliability and effectiveness of the assessment. In both cases, the Assessor had low confidence in the assessment results based on the acknowledgement that the re-interpretation was not robust (UTL) and the lack of clarity about the effect of an unticked box on the resulting, disappointing, score (ASA). Table 17 summarizes the highlighted characteristics and relates them to principles.

Table 17. Interpreting criteria and terminology is found challenging for the DPCMM

<table>
<thead>
<tr>
<th>Characteristics of the model</th>
<th>Relevant principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPCMM claims to be flexible, but no advice is provided on how to adapt or re-interpret criteria. Terms such as &quot;electronic records&quot; and “retention schedules” are used in DPCMM in a document targeted (among others) at libraries and companies.</td>
<td>1.1a) application domain 1.3 central constructs of the application domain 2.2c) Advice on adaptation and configuration of criteria</td>
</tr>
</tbody>
</table>

A similar set of principles come into play quite differently in the assessment projects using AIDA. When applying AIDA at ASA, the Assessor wondered what “highly integrated” means in the first technological infrastructure dimension, why a link to “external services” would be important, and how “national and international” would be relevant. The progression from stage 4 to 5 was unclear. The guidance to go with the lower level avoided the risk of false high rating, but reduced satisfaction with the assessment results. The assessor’s difficulties in mapping the terms to his environment negatively affected his confidence in the results.

This difficulty in interpreting criteria consistently corresponds to reports by others having applied AIDA who emphasized that the description of the criteria leaves “too much room to interpret what is supposed to be evaluated” and suggest that “With so much freedom to interpret many of the elements, the team can easily become consumed with how to apply the toolkit” (Miller, Blake, & Sorsby, 2012, p. 96). The characteristics and principles are summarized in Table 18.

Table 18. Interpreting criteria and terminology is found challenging for AIDA

<table>
<thead>
<tr>
<th>Characteristics of the model</th>
<th>Relevant principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDA does not fully define its terms and does not come with a glossary. The concept of externalization is applied throughout all aspects and legs. The toolkit encourages re-interpretation, but does not provide advice on how to do so. It contains advice to stay on the modest side in case of doubts.</td>
<td>1.2d underpinning theoretical foundations with respect to evolution &amp; change 1.3 central constructs of the application domain 2.2c) Advice on adaptation and configuration of criteria</td>
</tr>
</tbody>
</table>

In another set of events, the assessor at ASA struggled to reconcile inconsistent terminology when applying DPCMM: The term "Producers" is used consistently according to the OAIS definition, whereas "Users" is not. The glossary includes an entry for "Producers" that references the OAIS correctly, but not for "Users". It includes "Consumer" but the term is not used. Designated Community (7) employs the term “Users” instead of the standard OAIS terms Consumers. In Access (15), the meaning of "access to DIP" is not fully defined, but implies external access and equates "Users" with system users. The assessor could not rely on guidance whether the criteria should apply to repository software or organizational processes.

The lack of distinction between the ability to perform an activity and the actual activities performed (as defined in the original CMM) interacted with these terminology issues and led not only to a loss of efficiency (the assessor needed to figure out interpretation and mapping to context) but also
effectiveness: the lack of adaptation clarity led to existing gaps that were not identified, and thus a loss of usefulness. This is summarized in Table 19.

<table>
<thead>
<tr>
<th>Characteristics of the model</th>
<th>Relevant principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPCMM does not distinguish between the ability to perform and the actual activities performed (as defined in the original CMM).</td>
<td>1.1d) class of entities under investigation</td>
</tr>
<tr>
<td>There are incomplete and inconsistent definitions of concepts such as Users, Producers and Consumers.</td>
<td>1.2c) granularity of maturation</td>
</tr>
<tr>
<td></td>
<td>1.3 definition of constructs related to the application domain</td>
</tr>
<tr>
<td></td>
<td>2.2c) advice on adaptation</td>
</tr>
</tbody>
</table>

### 11.3 Conceptions of Maturity

In the cases applying DPCMM, the Assessors disagreed with how the model defined concepts and metrics of maturity. DPCMM defines its highest level of maturity as "optimal", not "optimizing", in contradiction to the CMM on which it claims to build. In effect, its articulation of levels and thresholds focuses on what the organization does, not on key aspects of maturity including how well the organization knows itself, how it measures its operations, or how it learns over time. The model focuses on capability, with little regard for learning, optimization, and other dimensions of organizational maturity, but uses the term “maturity” to describe this. The scales are not discussed or justified.

At ASA, the Assessor had significant prior exposure to notions of maturity in organizational management. He emphasized that none of the criteria address the degree to which processes are understood across the organization. In several components he had difficulty reconciling high levels in the DPCMM with his conceptions of high maturity and the desired levels of maturity in his organization. This made him lose confidence in the model. Both Assessors had difficulties with the component Collaborative Engagement (4) that defines a formalized perspective of collaboration. At UTL, the Assessor noted that important collaborations exist but may not be measured or captured in a formal document, because that may not be appropriate for these relationships at this time. At ASA, the Assessor noted that the required formality of a “collaborative framework” did not capture the types of collaboration taking place at his organization, which are informal but effective. Concerns were raised in both cases by the lack of clarity about how the entities under investigation change in the context of the application domain and by the requirement for collaboration to be formalized seemed inappropriate. This led to a loss of confidence and usefulness of the model. For less experienced Assessors, it could lead to significant risks of unsubstantiated overconfidence in their organization’s maturity.

The characteristics of the model revealed in these events show how the inconsistent application of concepts of capability and maturity relates to multiple design principles: The key concerns of maturity need to build on clearly defined concepts from the domain in order to provide an effective perspective in practice. This confirms the relevance of the findings from the analytic evaluation. Table 20 summarizes the relevant characteristics and principles.

<table>
<thead>
<tr>
<th>Characteristics of the model</th>
<th>Relevant principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPCMM uses the term “maturity” but uses criteria and threshold that measure domain-specific actions to be carried out. It effectively focuses on increased</td>
<td>1.1d) class of entities under investigation</td>
</tr>
</tbody>
</table>
capability and leaves important aspects of maturity unaddressed. The model requires collaborative engagement to be based on formal contractual arrangements.

1.2b) maturity levels and maturation paths
1.2d) underpinning theoretical foundations with respect to evolution & change

Another example illustrates how in both assessments using AIDA, the concepts of maturity based on the Five Stages (acknowledge, act, consolidate, institutionalize and externalize) led to incongruent results and dissatisfaction.

In both cases, the Assessors had difficulties to interpret the meaning of "externalization" in many of the dimensions of AIDA or simply found it inapplicable. This was combined with difficulty to interpret the distinction between Institution and Department, and gaps in criteria describing level 5. One Assessor in particular had difficulties reconciling his understanding with the definitions of the highest level. In many dimensions, AIDA’s definition of level 5 as externalization to him did not seem “better” than level 4. At UTL, the Assessor found it problematic that the use of externalize as the highest level places the emphasis on actions or processes that extend beyond institutional boundaries.

Assessors dealt with this inconsistently: They sometimes marked their organization down to the lower level to be on the safe side, as suggested in the guidelines. However, at ASA, the Assessor chose to apply the highest level in Sustainability of Funding (Resources 6) despite not meeting lower level criteria, because the involvement in externally funded project corresponds to Stage 5. At UTL, the Assessor noted the inconsistency in what Stage 5 is measuring and noted that the different ways of externalizing “all mean different things”.

Both Assessors mentioned that there is no separate mission statement for their department, unlike suggested by AIDA’s criteria. Both found that externalizing operations may not be appropriate for all organizations and that presenting this as ideal limits the framework’s usefulness for institutions that have justifiable reasons for limiting the involvement of outside agents in their preservation processes. As a result of this definition of the fifth stage, the framework failed to capture self-reflective practices of continuous improvement or to recognize these as an indicator of maturity. Efficiency suffered from the need to review and interpret criteria and map them to the organizational context. Effectiveness was affected by the incomplete mapping to the context and the associated loss of confidence.

The characteristics of the model that surface in these events highlight the importance of design principles that address the underpinning foundations of maturity and growth. Note that these are the same principles as in the previous set of events, as summarized in Table 21.

<table>
<thead>
<tr>
<th>Characteristics of the model</th>
<th>Relevant principles</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDA indiscriminately applies all elements to both levels of institution and department, and directly applies the ideal of &quot;externalization&quot; to both levels and all elements. In many dimensions, no criteria are given for the highest level.</td>
<td>1.1d) class of entities under investigation</td>
</tr>
<tr>
<td></td>
<td>1.2b) maturity levels and maturation paths</td>
</tr>
<tr>
<td></td>
<td>1.2d) underpinning theoretical foundations with respect to evolution &amp; change</td>
</tr>
</tbody>
</table>

Table 21. AIDA’s focus on Externalization is found problematic.
12 Conclusions

This report describes an evaluation method for assessment frameworks that takes into account the process of development, the design of a framework, its quality in use in the assessment process and how these influence outcomes. The comparative multi-case study considers organizational context and the fit of the framework with the organization. An approach based on minimal interference also helps to provide the best possible value to the practitioner conducting the assessment.

The findings discussed above highlight the fact that even though quality in use is not a clear-cut question, it can be investigated systematically. The combination of empirical and analytic evaluation supports us in explaining how a framework’s characteristics influence its quality in use. Ideally, for a framework to be perfectly suitable in a range of environments, a number of conditions should be met: The domain constructs that are used in its conceptual model should match the domain of the organization and be expressed in terms well understood by the organization’s stakeholders; the purpose of the model should contribute to the goals of the organization; and the entities under investigation should correspond to elements in the organization and be recognizable as such. Where perfect correspondence of domain concepts and organizational concerns is not given, the framework would ideally be flexible enough to reinterpret, situate, and adapt the model and its components. This requires robust terminology to ensure that correspondence of concepts can be established across communities of practice. Consideration of these aspects has led to the articulation of design principles for maturity models (Pöppelbuß & Röglinger, 2011).

The evaluated frameworks could be improved by revisiting these central constructs used in the design of maturity models; augmenting the documentation and guidance made available to support the assessment process; and reviewing central terms and definitions. However, to fully understand the causes of the quality in use issues and identify ways to address the concerns, an in-depth analysis of the relationship between the framework in its usage context, the model as a designed product, and the process contributing to the model’s design needs to be conducted. This is the focus of further analysis beyond this report, for which the summary of evidence presented here provides the starting point.

12.1 Limitations

The evaluation method presented here is intended to be applicable to any assessment framework, but it will be most effective in the evaluation of frameworks that facilitate self-assessment for two reasons. First, it involves considerable effort in analysis in relationship to the assessment project, so for an extended audit involving many sessions and participants, the effort may become prohibitive. Second, the method does not focus on verifying the accuracy of the Assessors’ answers and the degree to which these answers can be verified.

Some limitations to the application of the case study method in the four cases merit discussion and should be considered for similar studies.

**Generalizability:** The organizations selected for assessment are typical locations for digital preservation practice: memory institutions that have preservation as a part of their mandate and strive for OAIS-compliance or certification as Trusted Digital Repositories. The advanced state of their general capabilities limits generalizability of the specific results and findings for organizations in other areas, in particular if they do not have an explicit mandate for curation or preservation or capabilities that are less well-developed. However, the same case study method should be applicable.

**Expertise of assessors:** The expertise of both Assessor and note-taker in these particular instances removed potential needs for clarification or confirmation of terms, concepts, or acronyms used in recording judgments and commentary and made communication between the participants very straightforward. Assessors with different expertise may take longer, require a higher degree of effort, or necessitate the use of additional resources. Their assessment projects may bring to light qualities of the frameworks that did not surface in the cases considered. Organizing the assessment as a group
activity focused on achieving consensus would increase the required effort significantly, but in return add value through the reflective process of building shared understanding.

**Sequence:** In our study, out of necessity, assessments with each framework were carried out in short succession at both organizations. The order in which the frameworks were used was chosen to minimize bias, and no influence of conducting successive assessments was noticed during analysis. The absence of inter-rater validation in the scoring process limits the scores’ reliability, however, and prior experience of using other assessment frameworks cannot be ruled out as explanatory factor in the Assessors’ impressions of the frameworks, processes and results.

**Interference:** During the assessments, the observers were conscious of not interfering with the assessor or influencing his reasoning or judgments about the criteria. With the role of the one author who acted as the assessment note-taker, interference only occurred on two occasions early in the process within the governance component of the DPCMM assessment at ASA. The research team was monitoring the process in real-time and hence could intervene as such interference became apparent. The note-taker’s influence was taken into consideration in data analysis.
13 Summary and Outlook

As the discipline continues to mature, so do organizations in their professional abilities. To guide progression, organizations leverage diagnostic tools such as maturity models and other assessment frameworks to structure the evaluation of where they stand and how they can improve. A continuous stream of models and tools has been developed over the past decade that supports them in this endeavor. However, few empirical studies are available to the practitioner interested in choosing a framework, and to the researchers and framework developers interested in an evaluation of such frameworks. To develop its knowledge base, digital curation needs shared evidence and rigorous analysis.

This report has outlined an approach to empirical research that evaluates assessment frameworks in digital curation through systematic studies that provide evidence of what happens in the practice of assessment and enable the effective analysis of such frameworks. A method for case study research provides the needed focus, scope and consideration of multiple sources of evidence that facilitates the process of evaluating frameworks in real contexts. The collaborative application of the method by a team of practitioners and researchers to a set of frameworks across organizational contexts strengthens the insights that can be gained. Using multiple cases, where each framework is applied in two settings, supports comparison across cases and triangulation of emerging issues via multiple perspectives to increase the robustness of findings. The method was documented and illustrated in detail to facilitate replication in other contexts and to other frameworks.

The report thus aims to provide insights to practitioners, frameworks developers, and researchers. It aims to help practitioners to evaluate individual frameworks; support developers of frameworks by identifying areas of improvement and by providing a method of evaluation; and provide a model of empirical research that can be adopted, replicated, and extended by researchers in this domain. Our analysis offers insights into the frameworks studied and the method applied for studying them, and raises questions for further investigation.

The findings suggest that opportunities exist to improve the design and use of assessment frameworks in this field. The process of assessment itself was seen as more valuable than any particular model, and a number of issues raised with these models suggest that they can be improved. However, the findings also reflect the nature of assessment as a complex enterprise and raise as many questions as they answer. This report focuses on the method and its key findings from all three perspectives to support an analysis of the frameworks studied here in order to explain the issues uncovered and extract insights for the design of future maturity models in digital curation.

The open issues raised here point to the need for more in-depth investigation and study of the insights into quality in use, how these insights can be explained, and how they could be addressed. Such an understanding would benefit researchers and framework developers by providing a more thorough explanation of the design principles and methods for assessment frameworks, and practitioners, by supporting the improved design of such frameworks.

We have presented here not only the findings of the project to study frameworks in use, but also the case study method developed in the hopes that it will be adopted by others to contribute to a broader evidence base for how different frameworks can be applied in different contexts. In addition, to provide an effective contribution of these insights to design iterations of these frameworks, and to the knowledge base of the discipline in designing and evaluating such frameworks, we must consider the empirical evidence in light of the design process of assessment frameworks and its decisions, and learn from other disciplines that have tackled these questions before. Above all, research on the design of maturity models in disciplines at the intersection of Information Systems, Information Technology and Management Science have investigated these issues for some time (Maier et al., 2012; Wendler, 2012).
References


Appendix A: Pre-Assessment Survey

In your opinion, what are the objectives of digital preservation?

2. How would you define the following concepts:
   a. authenticity
   b. integrity
   c. designated community
   d. obsolescence

3. How do you intend to make use of the results of the assessment?

4. What do you know about the model to be used in the assessment? How do you know it?

5. Do you have any experience with assessing an organization’s ability for digital preservation?

6. To the best of your knowledge, has the digital repository ever conducted an assessment before?

7. Based on your current knowledge of the repository, what do you think of its current abilities for digital preservation?
   a. On a scale of 0 to 5 - where 0 is “non-existent” and 5 is “perfect” - where would you rate your repository’s ability to fulfill the purpose of digital preservation?
   b. How did you arrive at that rating?

8. Can you tell us about specific strengths in your repository’s digital preservation operations?

9. Can you think of specific weaknesses in your repository’s digital preservation operations?

10. How much do you agree with the following statements? (using a scale of: strongly disagree / disagree / agree / strongly agree)
    a. Assessments of my repository are essential for my continued confidence in our ability to preserve digital objects.
    b. I trust results of such assessments more if the models used for the assessment have been rigorously validated.
    c. I want to be able to conduct an assessment myself, without costly external support.

11. Do you anticipate taking any actions based on the outcomes of the assessment?

12. Will the outcomes be used:
    a. for improvement (Y/N)
    b. shown to other stakeholders in your organization (Y/N)
    c. shown to others outside (Y/N)
    d. as an input for working towards certification (Y/N)
Appendix B: Post-Assessment Interview Questions

I - PRELIMINARIES

1. What is your overall impression of the framework that you used?
2. Do you feel that the framework provided you with useful input?
3. Do you feel that the outcomes accurately represent the capabilities of your organization for digital preservation?
4. How did you make this judgment?

II - CLARITY, USABILITY, LEARNABILITY, APPLICABILITY

5. Did you find the framework and its concepts, structure, and tools easy to understand?
6. Were there specific issues?
7. How consistent is the framework with your understanding of the concepts, concerns, and principles of digital preservation?
8. Were there parts of the model that did not seem to apply to your organization?
9. Did you find the progression of the criteria to be logical? Did you notice any inconsistencies in how it progressed?

III - GUIDANCE

10. Did you find the guidance provided on the application of the framework to be sufficient and helpful?
11. Did you find the guidance provided on the interpretation of the results to be sufficient and helpful?

IV - EFFORT AND DURATION

12. What factors contributed to the time it took to complete the assessment?
13. Can you imagine a change in the team that would have made the assessment easier, faster or otherwise more efficient? Which?

V - OBSTACLES

14. What obstacles did you face in conducting the assessment?
15. How did you deal with these obstacles?

VI - RESULTS

16. On a scale of 1 to 4, how confident are you in the accuracy of the results of the assessment?
17. Would you be comfortable to ….
18. share it with your colleagues within the organization saying “This is an accurate diagnosis”
19. share it with your superiors within the organization saying “This is an accurate diagnosis”
20. make a public statement saying “This is an accurate diagnosis”
21. bet your job on it saying “This is an accurate diagnosis”
22. What (specific) concerns do you have regarding the accuracy of the results (if any)?

VII - FINAL THOUGHTS

23. Would you recommend other organizations like yours to do this assessment?
24. What’s the real value for your organization from this result?
25. Do you intend to do something with/about the result?
26. Is it relevant to you that the criteria and their evaluation are not backed up by evidence at this point, after you completed the survey?
27. Do you have any final thoughts or comments on the experience?