Challenges in preservation (planning)

Christoph Becker[1]

Vienna University of Technology, Vienna, Austria
http://www.ifs.tuwien.ac.at/~becker

Abstract. This short paper attempts to highlight some challenges to be tackled by DP research in the next years, taking as a starting point the perspective of preservation planning. These challenges are in short: (1) Scalability (up and down) requiring (2) measurement of relevant decision factors, in turn requiring (3) benchmarking and ground truth. (4) Quality-aware emulation. (5) Move from the current closed-systems approach to open structures that accomodate evolving knowledge. (6) Move from post-obsolescence actions to *longevity engineering*.

1 Introduction

This paper tries to highlight in a rather informal way some issues in digital preservation research that seem to pose particularly strong challenges, or problems for which effective solutions would have a strong impact on the way that operations are carried out in reality. A natural consequence is that the viewpoint taken here strongly stems from the work I have conducted over the past years in preservation planning [1, 3], component selection [4], and quality-aware migration [2, 5].

Taking this perspective, where are we now? Over the past years, considerable effort has been invested in analysing the factors contributing to decision making and the constraints posed by different scenarios, and in building decision-making frameworks and tools. With current state-of-the-art procedures in digital preservation, we can define organisational constraints, and we can create plans that treat a certain part of a large repository. The Planets preservation planning methodology defines a structured workflow for creating preservation plans. The planning tool Plato developed within Planets follows this well-established workflow to build trustworthy preservation plans[1]. The tool produces substantial evidence as documentation underlying the decisionmaking procedures, and has experienced significant uptake in the DP community.

A resulting plan is currently able to define treatment for a well-defined set of objects. It is constructed largely manually, albeit tool support is increasing; it normally is not applicable to heterogeneous holdings; it is not deployed and executed automatically in a repository; and all plans need to be monitored manually. Further, no mechanism exists today to relate preservation policies directly

[1] http://www.ifs.tuwien.ac.at/dp/plato

Dagstuhl Seminar Proceedings 10291
Automation in Digital Preservation
http://drops.dagstuhl.de/opus/volltexte/2010/2768

to preservation plans. So far, the high level policy goals have to be correlated intellectually (manually) to specific preservation goals.

Preservation planning takes place in a context of evolving technologies, user communities, and organisational policies. This triage defines the constraints in which decision making needs to operate. Some of the automation issues can be addressed in a relatively straightforward manner in the next years; but some depend on more complicated challenges.

2 Challenges

2.1 Scalability One: Down

While several large-scale approaches to digital preservation have been fairly successful, smaller institutions and individuals have not yet been able to take advantage of these methods and tools. Yet, a large amount of information, comprising an enormous value, is created and stored every day by private users and small organisations. This ranges from family photographs and videos to emails and other types of documents created in virtually every home and office. Small and medium enterprises face similar challenges concerning their core business documents. These objects need to be preserved through solutions with low entry barriers, affordable running costs and clearly communicated benefits [9].

2.2 Scalability Two: Up

Emerging applications of grid and cloud technologies promise to deliver scalable operations for repositories and preservation actions. But fundamentally, for a system to be truly operational on a large scale, all components involved need to scale up. We need an approach to planning, monitoring, and operating a repository on a petabyte-scale. Only scalable monitoring and decision making enables automated, large-scale systems operation by scaling up the decision making and QA structures, policies, processes, and procedures for monitoring and action.

Increasing automation in decision processes such as preservation planning will include the following aspects.

- Automated selection of representative sample content based on large-scale in-depth collection profiling;
- Automated construction of criteria trees with a certain coverage of influence factors, based on formalised policy models that reflect environmental and organisational constraints;
- Automated construction of significant property trees based on a combination of templates and properties extracted from sample objects;
- Automated construction of utility functions based on measured values, policies, and (aggregated) user feedback; and
- Automated suggestion of candidate components to evaluate, based on shared experience bases and aggregated utility values.

2.3 Measurements

The goals of scalability require reliable, repeatable and efficient *measurement* of the decision factors that underly all DP and PP operations, such as

- Desired properties of digital objects,
- Properties of digital objects that have to be kept through changing representations and evironments,
- Properties of formats and other representation information networks, and
- Operational properties of systems and components.

Some work has addressed the second aspect [5, 6]. Practical experience indicates that instead of fundamentally canonical approaches to ensuring authenticity, which encounter tough challenges hidden in the small but abundant complexities and variations of format implementations, a more pragmatically viable way is to define a roadmap of aspects that need to be measured and address them on a prioritisation basis. This requires quantified impact assessment of decision factors and measurements to allow prioritisation, and it requires models and methods for addressing measurement reliability and uncertainty.

2.4 Benchmarks and ground truth

Any improvement on the coverage and precision of these measurements is doomed to fail if it cannot rely on substantial benchmarks and well-known ground truth that supports validation.

The way currently available QA mechanisms are developed to support, for instance, migration processes, resembles throw-away prototypes created without formal requirements specifications. It is an exploratory way of investigating potential paths rather than systematic improvement. When developing a mechanism to measure property x of a conversion process, we need a way to judge whether it is an improvement, i.e. to verify how this mechanisms' measurements differ from others and whether they are correct or not. This rather obvious statement, unfortunately, requires a substantial body of benchmark data with annotated ground truth. DP benchmarks have been discussed earlier [7], but no well-defined benchmark exists so far, since the creation of such a corpus requires substantial resources.

One reason is the black-box view that we generally take on defining this ground truth: We analyse collections of digital objects *ex-post* and try to figure out what they contain. This is obviously susceptive to the very same problem we intend to improve upon: Measurement uncertainty and lack of coverage.

2.5 Quality-aware emulation

Despite the large gaps, there is some progress on measuring quality of migration processes. When relying on other approaches such emulation or virtualisation, however, there is a fundamental lack of solid QA approaches. If emulators cannot be verified or tested for quality, it becomes very hard to justify a decision for

adopting a particular emulation technique or toolset as a preservation strategy. Automated mechanisms are needed to measure operations in emulation environments. This problem is of considerable complexity, as illustrated by a related recent discussion [8].

2.6 Closed systems, open knowledge

The majority of current DP efforts aim to build systems to solve certain problems. We build repositories, migration engines, planning tools, QA workflows, and format registries. Designs for these systems mostly rely on inherently closed-world models. Moderated registries such as PRONOM that are in use today are not dynamic enough to capture the evolving facts and the knowledge that is available, for example on the web. Obviously, the implicit closed-world assumption in the design of these systems does not hold in reality. For registries, open information models using RDF and ontologies have to be leveraged to capture the inherently evolving nature of repositories, user communities, and technologies, and allow reasoning over known facts to produce derived knowledge.

2.7 Longevity engineering

The general trigger for a preservation activity today is impending or acute obsolescence of a certain object or set of objects, independent of whether the object is a document, a data set, or a software system. We can interpret this obsolescence as a *fault* caused by a *bug* in software in turn caused by an error committed when creating the software, probably due to a (sometimes unaviodable) failure to acknowledge a future development. We can also see it as a case of necessary *adaptation* of a software system in the course of systems maintenance. In both cases, it is well known that the cost of changing software rise continuously, sometimes exponentially, the later errors are detected in the development lifecycle. This is obvious for software engineering and has spurred developments such as prototyping, agile methods, test-driven development, and test-driven design. However, in digital preservation we are still mostly acting on an ex-post and ad-hoc basis instead of building longevity into our digital artefacts from the start. Establishing *longevity as a fundamental non-functional requirement* in software engineering from the start – and finding the right tools and design principles to address it the way security or availability are addressed – is currently a far goal. But it would potentially have a profound impact on the resulting longevity of the systems we build and rely upon.

3 Conclusion and Outlook

Digital preservation needs to move from one-off decision-making and ex-post activities into a continuous, and continuously optimising, management activity. This requires reliable and efficient measurement of operations, which in turn will not be achievable without well-established ground truth and benchmarks.

Measurement reliability and efficiency need to be modelled and addressed; emulation needs to become quality-aware; and models need to become more open to discovery of evolving knowledge.

Finally, instead of conducting quasi post-mortem actions on obsolete objects, instead of merely reacting to obsolescence, we should aim to introduce longevity as a fundamental non-functional requirement and design principle in software engineering.

References

1. Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, and Hans Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries (IJDL)*, December 2009. `http://dx.doi.org/10.1007/s00799-009-0057-1`.
2. Christoph Becker, Hannes Kulovits, Michael Kraxner, Riccardo Gottardi, Andreas Rauber, and Randolph Welte. Adding quality-awareness to evaluate migration web-services and remote emulation for digital preservation. In Maristella Agosti, Jose Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas, editors, *Research and Advanced Technology for Digital Libraries. Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009)*, volume 5714 of *LNCS*, pages 39–50. Springer, September 2009.
3. Christoph Becker, Hannes Kulovits, Andreas Rauber, and Hans Hofman. Plato: A service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL 2008)*, 2008.
4. Christoph Becker and Andreas Rauber. Improving component selection and monitoring with controlled experimentation and automated measurements. *Information and Software Technology*, 52(6):641–655, June 2010.
5. Christoph Becker, Andreas Rauber, Volker Heydegger, Jan Schnasse, and Manfred Thaller. Systematic characterisation of objects in digital preservation: The extensible characterisation languages. *Journal of Universal Computer Science*, 14(18):2936–2952, 2008. `http://www.jucs.org/jucs_14_18/systematic_characterisation_of_objects`.
6. Michael Hartle, Arsene Botchak, Daniel Schumann, and Max Mühlhäuser. A Logic-based Approach to the Formal Description of Data Formats. In *Proceedings of The Fifth International Conference on Preservation of Digital Objects (iPRES)*, pages 292–299, London, United Kingdom, September 2008. The British Library.
7. Robert Neumayer, Christoph Becker, Thomas Lidy, Andreas Rauber, Eleonora Nicchiarelli, Manfred Thaller, Michael Day, Hans Hofman, and Seamus Ross. Development of an open testbed digital object corpus. DELOS Digital Preservation Cluster, Task 6.9, March 2007.
8. George Phillips. Simplicity betrayed. *Commun. ACM*, 53(6):52–58, 2010.
9. Andreas Rauber, Christoph Becker, Hannes Kulovits, Michael Greifeneder, Petar Petrov, and Stephan Strodl. Digital preservation: From large-scale institutions via SMEs to individual users. In *International Conference on Digital Libraries (ICDL 2010)*, New Delhi, India, February 2010.