(Semi-)Automated digital preservation archives for small institutions and private users

Stephan Strodl
Vienna University of Technology
Vienna, Austria
strodl@ifs.tuwien.ac.at

Introduction

Large heritage institutions have been addressing the demands posed by digital preservation needs for some time. In contrast small institutions and private users are less prepared to handle these challenges. An increasing quantity of digital collections is held by small institution with limited know-how and awareness of digital preservation. Digital assets are becoming more important for an increasing number of institutions in the long run (e.g. legal obligation, intellectual property or business data). The limited resource in these institutions for archiving drives the need for new approaches of (fully or semi)-automated archiving systems. Research and development in the area of digital preservation is mainly done by memory institutions and large businesses. Consequently, the available tools, services and models are developed to meet the demands of professional environments.

Automated archiving systems are needed for institutions with little professional know how in digital preservation. Important aspects are hiding the complexity of the processes, providing support for decision making and automated error handling. The automation of preservation workflows raises a number of research questions, e.g. metadata management, quality assurance and tolerable limit of loss of preservation actions and automated preservation planning.

In this paper we will illustrate some research challenges in the field automated preservation. Current initiatives and research activities are indicated that are tackling aspects of the research topics.

Automation

In order to enable small institutions to manage and preserve their digital assets, the complexity of digital preservation has to be reduced based on established best practice examples. Simple and automated services are required with special support for decision making and error handling. These

Dagstuhl Seminar Proceedings 10291
Automation in Digital Preservation
http://drops.dagstuhl.de/opus/volltexte/2010/2764

services must have small entry barriers for new users and should presuppose little if any knowledge of the digital preservation domain. The automation comprises almost all functions of an digital preservation archive and it requires new methodology approaches as well as enhanced tool support for preservation tasks. In terms of digital preservation critical automatons are in particular preservation planning, preservation action and the quality assurance of the preservation results.

The Hoppla [6] archiving system provides a digital preservation solution specifically for small institutions. It combines back-up and fully automated migration services for data collections. The system allows user-friendly handling of services and outsources digital preservation expertise via an update service. Migrations plans and services are provided by an external update service to the client side of the archive.

Other projects provided automation for specific preservation task. The CRiB project [5] provides an intelligent decision support for migrations. CRiB is a Service Oriented Architecture (SOA) providing recommendation for migrations and server to carry out format migrations. The recommendations are baed on performance measurements suitability of formats and data loss during migrations. The CRiB service was integrated into the RODA archive [4]. The Protage[1] project is exploring the potential of interaction between intelligent software agents and Web services for automating digital preservation tasks.

Beside the predominant preservation strategy migration work on automation of emulation is done in the KEEP research project [2]. Developing an emulation framework supporting the virtualisation of different emulators.

Policies, Requirements, Usage

Policies have a clinical importance in automated repositories as they form the basis of decisions making. The requirements, needs and preferences of the archive users and institution need to be formalised in order to process them by the software. Further information that needs to be formalised includes technical aspects of the archive and information about the collection.

The design of models for policies is a trade off between generalisation and level of detail. On the one hand the model should be applicable to all kinds of institutions, collection and settings. On the other hand the model should cover and represent the individual requirements and characteristics of each setting.

Other difficulties of polices are portability, system independencies and multiple policies. Usually we have different levels of policies, system-wide that are usually very generic and general down to more detailed ones about

[1] `http://www.protage.eu`

specific collections. Intelligent frameworks are needed to manage and integrate the different policies levels and break the information down for individual applications. The multiple policy levels always bear the risks of contradictions and overlaps. The challenge of the archival software is the breakdown of generic formulated terms of the policies to particular decisions and actions in the archive.

Moreover policies should be system independent and portable that they can be used institutional wide in different application. A basic requirement for such a policy framework is a common vocabulary and understanding of the terms. Ontology as used in information science can be a promising approach a common policy framework.

Standards, certification, best practices

The confidence of the consumers is a core aspects of digital preservation system. At present the vendors are claiming their digital preservation system compatible to a variety of reference system, recommendations and standards form other disciplines.

Criteria and standards for digital preservation repositories are an active field in the community. Examples are the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) [7] and Catalogue of Criteria for Trusted Digital Repositories from the certification working group of NESTOR [3].

In the community exists a common understanding about the high level duties and responsibilities of a digital preservation system (based on OAIS model [1] and other de-facto standards). What is currently missing is a breakdown of the high level goals into measurable evaluation criteria with well defined minimum requirements. This could build the basis for a standardized certification inactive for digital preservation repositories.

For automated small preservation repositories more specific standardised evaluation can be a practical approach. It could include the evaluation for specific settings for example a specific collection type (images, office documents) or user type (photographer, architect).

Another important aspect in terms of certification is legal obligations of companies for digital preservation. Well defined legal regulation and specification for digital preservation repositories are currently missing.

Slow progress is made in the community of digital preservation in collection and consolidating best practices examples. Collection of best/good practices of digital preservation implementation could provide a entering guide for people that are interested in DP. An example for best practice collection is the IT Infrastructure Library (ITIL)[2] for IT information systems.

[2] `http://www.itil-officialsite.com`

Complex objects

Most current research and work in digital preservation focus on singular digital object type (e.g. text and image documents). Thus ignore the fact that most of the digital objects have explicit or implicit dependencies to other objects. Preserving the dependencies of digital objects is an important aspect for the future use and access of the data. Prominent examples for complex objects are HTML-pages, videos container formats, e-mails, presentation with embedded multimedia objects.

We have different kind of dependencies between objects, e.g. contextual, technical. Example for a contextual relationship is a single image of a beach, only in context of a collection (e.g. holiday 2009, bahamas) the image becomes valuable for the user. Technical relationships can be embedded (such as videos in presentations) or used (specific codec for videos or plugging for web-page context). There are different approaches and models to specify and classify relationships between objects.

For automated preservation archives a pressing challenge is the automated detection of relationships between objects. The technical relationships have to consider for preservation actions.

Storage Utilisation, Forgetting, Deleting, supported Access

Today the amount of data grows rapidly every year. Even with increasing storage capacity and cheaper storage media the efficient usage of the storage is a requirement for all archives. New approaches are required to help users with appraisal and also deleting objects of the archive. The determination of future value of digital assets is a difficult task. Deliberate appraised objects in an archive increase the future value and improves the use of the archive. An approach to support the appraisal can be for example to capture usage statistics of digital objects on source system before archiving them. Another approach can be the reappraisal of archived objects after a specific period of time. Work on appraisal and disposal for archive was done by PARADIGM project [3].

Cognitive techniques, particularly forgetting, are interdisciplinary research topics that can support various preservation tasks. The amount of data stored in an archive can grow very quickly and forgetting can be used for advance storage utilisation. New approaches can help to determine objects (or version of objects) that are more preferable to forget or aggregate (e.g. preserved videos or photos in a lower resolution).

[3] `http://www.paradigm.ac.uk`

Conclusion

In this article we pointed out some research issues for automated digital preservation archives. The automation of the processes raises the number of requirements for software developments of modules. The need of common approaches has been identified for different open issues. In this context a fundamental requirement is the definition of a ground truth for digital preservation settings. Common knowledge bases and defined best practice can form the basis for enhances preservation approaches and wide accepted standards.

Other research disciplines can be vital source of inspiration for new application area of existing techniques and new approaches for digital preservation. The here presented list of challenges is not complete and they are not fully analysed. It should provide a rough guide for thoughts to encourage consideration.

References

[1] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[2] KEEP Consortium. KEEP - keeping emulation environments portable. `http://www.keep-project.eu`.

[3] nestor Working Group -Trusted Repositories Certification. Catalogue of Criteria for Trusted Digital Repositories. Tech. rep., nestor - Network of Expertise in long-term STORage, Frankfurt am Main, June 2006. Version 1.

[4] Portugese National Archives. Roda - repository of authentic digtial objects. `http://roda.di.uminho.pt`.

[5] Ramalho, J. C., Ferreira, M., Faria, L., Castro, R., Barbedo, F., and Corujo, L. Roda and crib a service-oriented digital repository. In *Proceedings of The Fifth International Conference on Preservation of Digital Object (IPRES 2009)* (2009), The British Library.

[6] Strodl, S., Motlik, F., Stadler, K., and Rauber, A. Personal & SOHO archiving. In *Proceedings of the 8th ACM IEEE Joint Jonference on Digital Libraries (JCDL'08)* (Pittsburgh PA, USA, 2008), ACM, pp. 115–123.

[7] The Center for Research Libraries (CRL), and Online Computer Library Center, Inc.(OCLC). Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). Tech. Rep. 1.0, CRL and OCLC, February 2007.