Collaborative Preservation Infrastructure - A Research Challenge

Rainer Schmidt
AIT Austrian Institute of Technology
Donau-City-Strasse 1, Vienna, Austria
firstname.lastname@ait.ac.at

ABSTRACT

Digital preservation environments like repositories and archiving systems are traditionally designed to operate autonomously and self-contained. We argue that a paradigm shift towards collaborative preservation environments can greatly improve technical as well as economical factors, and provide major benefits to the user community. The research challenge described here envisions improvements to state-of-the-art preservation systems based on three design goals: (1) collaborative storage and backup, (2) participatory data analysis, and (3) coordinated sharing and outsourcing of hosting infrastructure. This article briefly describes the motivation for the research challenge and explains the problem it is aiming to solve. Furthermore, initial design considerations and pointers to existing approaches are sketched.

Keywords

Research Challenges, Digital Preservation, Collaborative Environments, Scalability, IT-Infrastructures

1. INTRODUCTION

Digital preservation is dealing with the problem of maintaining digitally encoded information so that it remains available and understandable over very long periods of time. While having its origin in the cultural heritage domain, digital preservation targets almost all areas of society and forms of digital information. Applying a suitable preservation strategy for individual (groups of) digital items can be a technically challenging task. The required effort typically depends on the individual objects and characteristics like hardware and software dependencies, data formats, or the integrity of the object, to name a few. However, in recent years, preservation strategies and corresponding technologies have been developed that enable us to preserve a large range of digital materials. Key aspects target the prevention of data loss (e.g. using replicated storage and checksumming) as well as format/hardware obsolescence (e.g. through continuous format migration and software/hardware emulation).

Another increasingly difficult problem digital preservation is facing, is the management of the steadily growing volumes of data. Analogous to web data processing [1] and scientific data management systems [2], preservation environments will have to adopt *Big Data* platforms in order in order to cope with the volumes of data that are being produced by today's society. The SCAPE project [3] is presently investigating exactly this research question aiming at developing tools and services for the efficient planning and application of preservation strategies for large and heterogeneous data collections. The Preservation Platform [4] developed in this context, supports the development of scalable preservation environments in terms of computation and storage. We expect that solutions that are built on such large-scale data management frameworks, can greatly advance the capabilities of existing preservation systems with respect to robustness, throughput, and scalability. Although a scalable platform, will help individual institution in managing and preserving growing amounts of data, we believe that individually hosted stand-alone systems won't be technically as well as economically viable on the very long term. The research challenge we introduce in this paper, called *Collaborative Preservation Infrastructure*, motivates the development of preservation systems that interoperate across institutional boundaries. The goal is to enable federated and highly optimized usage of it-infrastructure in order to securely preserve data at large-scale.

2. MOTIVATION AND PROBLEM DESCRIPTION

We argue that in order to provide viable solutions on the long run, there is is a fundamental change required in the way preservation environments work. At present, most archival systems are designed as autonomous systems. These systems might be capable of managing data within in a distributed environment but hardly know about information that is stored within different systems in the same environment. This design philosophy however causes a number of inefficiencies which can have a significant impact on the overall scalability and cost efficiency of presrvation system. These factors are in particular critical if systems are operated within a large-scale environment. An example is the uncontrolled replication of content. One could consider multiple web archive which independently harvest overlapping content, which is stored in different environments, and additionally saved as a backup in different locations. From an overall perspective, such systems lead to an inefficient and expencive use of it-resources.

Truly collaborative environments would detect replications and - for example - act as mutual backup resources. However, archival storage and backup provides only one aspect of collaboration. Computation provides another major aspect. Preservation environments and the institutions that operate them are typically limited in the number of computations they run against archived data. Examples are regular checksumming, identification, and migration of archived data. We argue that it will be important to provide means that allow 3rd party users to process archived content because with a growing amount of content it will simply be impossible for a single institution to understand and curate all of the data it preserves. The major argument here is that there is no point in preservation data if it cannot be analyzed and, hence, be discovered by its users. The third aspect, this research challenge targets are limitations in the hosting model. Infrastructure providers that offer the hosting of applications and data within globally operating data centers (commonly called clouds) have gained major attention in recent years. The model has turned out to be very cost efficient taking advantage of the *economy of scales*. It is obvious that institutionally hosted data repositories can only be operated for very limited (hand picked) data sets without hitting an economic barriers. Cloud offerings, on the other hand will usually not meet institutional policies, hindering one to outsource the housing of the data and systems. The same policies will most likely also prevent collaborative storage and computation. We therefore see an important research challenge in the development of methods that foster and promote the development and deployment of collaborative, scaling, and globally operating preservation environments.

3. INITIAL DESIGN CONSIDERATIONS

Grid Computing is a discipline that deals with the coordinated sharing of computer reserouces over multiple administrative domains. While this concept - in contrast to the cloud model - did not prevail as a general purpose architecture, it has been very successful in providing collaborative infrastructures for large-scale data management and computation in specific domains. CERN's LHC Computing Grid provides one of the most prominent examples for that [1]. An interesting research question would be to study the feasibility of building a smaller but similar structured infrastructure for the purpose of digital preservation. The concept of tiers might, for example, be well applicable in building different *trust domains* of a preservation system. A limited number of highly sensitive data sets (a.g protected and provate data) could be preserved by its owner institution only, while large amounts of less sensitive data (e.g. scans of commercially printed books, newspapers, web harvests) could be hosted in less protected tiers (or circles) of the infrastructure. Such circles could for example include (1) an isolated stand-also tier, (2) a tier that is secured but connected and interoperating with other systems, (3) a tier that manages data using cloud resources. Other parameters that could be relaxed depending on the tier might be related to safety, security, and preservation policies the data is subject to. In general, we expect that collaboration will need strong means to control what and how resources are shared between the stakeholders. However, we also expect that collaborative preservation environments have the potential to be greatly beneficial for institutions that preserve data as well as for the user community.

[1] http://public.web.cern.ch/public/en/lhc/Computing-en.html

4. REFERENCES

[1] Dean, J., and Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM 51* (January 2008), 107–113.

[2] Hey, T., Tansley, S., and Tolle, K., Eds. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[3] King, R., Schmidt, R., Becker, C., and Schlarb, S. SCAPE: Big Data meets Digital Preservation. *ERCIM News 2012*, 89 (April 2012).

[4] Schmidt, R. An Architectural Overview of the SCAPE Preservation Platform. In *9th International Conference on Preservation of Digital Objects (in press)*. (Toronto, Canada, October 2012).