

Hoppla - Digital Preservation Support for Small Institutions

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

Florian Motlik
Vienna University of
Technology
Vienna, Austria
motlik@ifs.tuwien.ac.at

Andreas Rauber
Vienna University of
Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

Small businesses (small office/home office, SOHO) have tremendous amounts of digital information. At the same time, they have little to no expertise on how to manage it, not to mention caring for their long-term preservation, as even simple back-up strategies pose already drastic challenges.

This demo presents the Hoppla archiving system¹ to provide digital preservation solutions specifically for small institutions and offices. It hides the technical complexity of digital preservation challenges by providing automated services based on established best practice examples. Appropriate preservation strategies and required tools for the collection are delivered via a web service, effectively outsourcing the required digital preservation expertise.

General Terms

Digital Preservation, Long Term Access, Archiving, SOHO

1. HOPPLA CONCEPTS

Hoppla² - (Home and Office Painless Persistent Long-term Archiving) presents a new approach to automated digital preservation systems that are suited to their needs. Requirements for digital preservation of holdings in small office settings differ from those in professional settings caused by different levels of expertise and skills of the users, the different environments, and objectives. The requirements in automation of the archiving process are higher for an archiving software solution for users with less expertise in preserving and managing collections. In institutional settings, critical decisions in preservation endeavours can be made by skilled staff. Existing open source digital repositories,

¹Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

²<http://www.ifs.tuwien.ac.at/dp/hoppla>

such as Fedora³ and DSpace⁴, are useful environments for professional archiving, but usability and required knowledge for configuration and use do not meet the skills of user in small institutions and SOHOs. The requirements in data management as well as managing the preservation task fully automatically pose significant research challenges. Solving these will represent an important step forward in enabling non-experts to preserve their digital data in very much the same way and at acceptable levels of quality as they can currently preserve conventional objects for their own needs.

The underlying principle of the system is finding a best effort solution with respect to the available technology and skills of the users. We cannot assume a highly sophisticated computer environment; neither can we expect a profound knowledge in digital preservation or archiving. With Hoppla we are currently developing a solution that combines back-up and fully automated migration services for data collections in small institutions and SOHO settings. It will combine bit-stream preservation via LOCKSS-style back-up⁵ with logical preservation by automatically obtaining migration rules and tools. The system should provide the best available and most practical preservation solution.

The system builds on a service model similar to current firewall and antivirus software packages, providing a user-friendly handling of services, an automated update service and hides the technical complexity of the software. Data is ingested from a number of sources such as data carriers, email repositories and on-line storage locations, while back-up storage supports off-line and on-line storage media in both write-once as well as rewritable forms.

A detailed description of the concept of the Hoppla system is presented in [1]. This demo provides an insight into the current development of the system and challenges that an automated archiving system is dealing with.

Figure 1 shows the basic architecture of the Hoppla system that is influenced by the OAIS reference model (ISO 14721:2003). It consists of four core components: acquisition, ingest, data management, preservation management and storage management. Two registries contain preservation rules and services. Both registries are updated automatically by an external update web service. The *service*

³<http://www.fedora.info>

⁴<http://www.dspace.org>

⁵<http://www.lockss.org>

registry contains services and tools for object identification, characterisation, preservation, and preservation validation. The registry also contains representation information about formats, for example the format specification. The *preservation rule registry* specifies preservation strategies for different types of objects.

The archiving system supports the acquisition of data from different sources via an API for acquisition plugins. The use of plugins allows to support all kinds of storage media and current as well as future data sources. The first version of Hoppla supports the disc acquisition and e-mail accounts. The selection of the digital objects to be preserved is performed in the ingest component. The module uses DROID⁶ to determine the object's format. The ingest component creates a collection profile that is used to identify appropriate preservation rules for the collection.

The data management enriches the objects in the archive with metadata to ease later reuse. Metadata are created from the additional information captured by the acquisition component, the documentation of migration processes, and metadata extracted from objects. The Hoppla system is designed to deal with complex objects, objects that are represented by several files that are related to each other (e.g. web pages).

Preservation management controls the logical preservation of the objects. This means that it is responsible for performing migration strategies on the objects in its archive. To do this preservation tools and rules are requested from an update service. The web update service consists of two web services, a resource based registry and a dialog based service. The resource registry provides executable tools and services for migration and characterisation and metadata for object formats. The dialog based service provides preservation rules for the collection on the basis of functional requirements and constrains. The functional requirements include the collection profile and specific user requirements. The constrains for the preservation rules are primarily available storage and when indicated costs of preservation services or additional storage.

The application flow of Hoppla starts with the ingest of new objects from source media, the object's formats are identified and a collection profile is created. Based on the collection profile suitable preservation rules are recommended by the web update service. For privacy reasons, the user can define the level of detail of the collection profile that is provided to the web update service. This way the user has strict control, which and how much metadata is sent to the server. New preservation rules and services are downloaded from the web service and preservation actions are performed on the client side. The new objects including the resulting objects from preservation actions are ingested into the collection and stored on storage systems.

The access module provides services that allow users to access the data stored in the archive. The access module further displays information about object dependencies and versioning history.

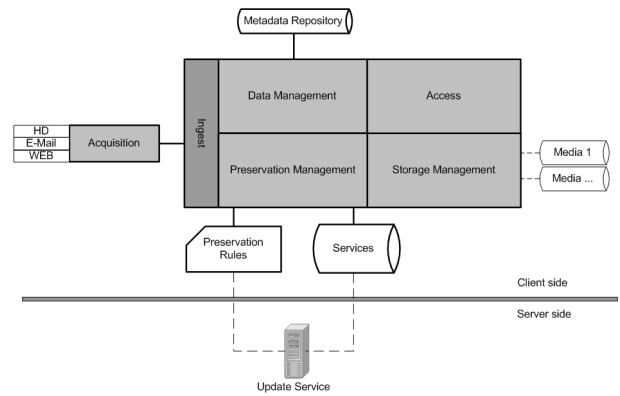


Figure 1: Architecture of the Hoppla System

Storage management is responsible for bitstream preservation. The data provided by the data management component are stored on various storage media. The storage management supports multiple copies of the data, following the concept of the LOCKSS project. In order to store the data on various storage systems or media, storage management implements a reduced version of a storage resource broker⁷. It provides a storage interface to access different storage systems, such as file systems or online storage system as well as write-once and rewritable media, by using plugins. All information and documentation generated by the Hoppla system are stored in XML format on the storage media.

2. RESEARCH ASPECTS

The first version of Hoppla focuses on basic migration, the web service for preservation rules and services, and the data management. A first version of the update service has been deployed providing preservation rules and services and an interface for administration. The functionality of the web update service will be further expanded and we specifically perform research on supporting the selection of preservation strategies considering functional requirements and constrains. Further improvement of the preservation rule selection based on the history of the collection will be done.

Another research aspect of the Hoppla system is context analysis of objects. The context of objects is essential for the interpretation of information objects and can provide enhanced access functionality to a collection. Various aspects of context in different dimensions can be automatically detected, and different views at multiple levels of granularity allow the extraction of the most appropriate connections to other digital objects. Information retrieval, data mining and OLAP-based approaches are used for structuring object collection and identifying relationships in multiple dimensions such a time, people, acronyms, content, etc.

3. REFERENCES

- [1] STRODL, S., MOTLIK, F., STADLER, K., AND RAUBER, A. Personal & SOHO Archiving. In *Proc. of the 8th ACM IEEE Joint Conf. on Digital Libraries (JCDL'08)* (Pittsburgh PA, USA, 2008), ACM, pp. 115–123.

⁶<http://droid.sourceforge.net>

⁷<http://www.sdsc.edu/srb/index.php>