PHAIDRA - A REPOSITORY-PROJECT OF THE UNIVERSITY OF VIENNA

Markus Höckner
University of Vienna
Dr.-Karl-Lueger-Ring 1, 1010 Wien

Paolo Budroni
University of Vienna
Dr.-Karl-Lueger-Ring 1, 1010 Wien

ABSTRACT

Phaidra (Permanent Hosting, Archiving and Indexing of Digital Resources and Assets) is used as a long-term preservation system through the assignment of persistent identifiers (permanent links). The project was launched in 2006 and is the successful result of cooperation between the Vienna University Computer Center, the Center for Teaching and Learning and the Vienna University Library. At present, Phaidra contains about 60,000 objects, all of which are provided with structured metadata.

1 INTRODUCTION

Like any other organization, the University of Vienna preserves millions of born-digital documents in several databases and file systems. Researchers are faced with two major problems: firstly, when they need to share their digital assets, they usually solve this problem by exposing their data on the internet (e.g. via a website). In this case, special knowledge is required to secure such precious assets in order to secure them from public downloading of these objects.

Secondly, researchers must provide a solution for long-term preservation [7]. Today, we produce millions of megabytes of data every day and we pay almost no attention to the fact that this data may not be accessible or reusable in the next few years.

Often the preferred solution is to build a repository. As a consequence, many repositories containing different technologies and metadata were built worldwide. But is there some method or solution available that would enable these to interoperate and exchange data?

In this context, the problem of long-term preservation is often underestimated. Only experts deal with this problem and they try to attract the attention of the producers of digital data regarding access problems which could occur in the future.

The University of Vienna has thus decided to address this challenge and started its own university-wide project in 2006.

2 NON TECHNICAL INNOVATIVE APPROACH AND SOLUTIONS

2.1 The long term perspective

Phaidra stands for a long-term archiving and digital asset management system and enables employees in teaching, research and administration to save, document and archive digital data and resources over a long period of time. Data can be systematically collected, equipped with multilingual metadata, assigned various rights and made accessible worldwide and around the clock. The continuous citability allows the exact location and retrieval of prepared digital objects. Phaidra can be actively used by all *staff and students* of the University of Vienna (via mailbox or u:net account). The objects can be viewed worldwide. Phaidra was developed at the Vienna University Computer Center in cooperation with the Vienna University Library and Archive services and the Center for Teaching and Learning. The project management is located in the University Library.

2.2 Project organisation and history

The project Phaidra bears three acting bodies: the Advisory Board, the Project Management and the so called Pilots.

The Advisory Board, an inter-university group of experts, performs strategic functions and supervises the achievement of strategic goals. The Project Management (two employees), based at the Library and at the Computer Center (three developers), is the operative entity of the project. The pilots for this were formed by clustering the needs of a few larger customer groups, like faculties, huge scientific projects as well as the University Library itself. The common challenge was to offer a vision of how to build a digital asset management system able to respond to the special demands of every department, institute and individual working at the university. The responses to this challenge were very important for the acceptance of the project by the potential user groups involved.

The project itself was approved for three years by the authority of the University of Vienna. Only after one year of developing the first release (April 2008) has been deployed. Two years later, in February 2010, version 2 of Phaidra has been presented.

In April 2010 the development of the project has been extended for another three years. One of the main reasons was that there is a big demand of such a system at the University of Vienna. A lot of projects and organizations show interest in such a system.

2.3 Relevant factors

Phaidra was expected to host all digital assets stemming from the fields of

- research,
- teaching,
- technology enhanced learning and
- management,

and therefore it now offers general search functionalities covering the full range of stored assets as well as faculty and project-specific areas.

Concerning the metadata, a part of the chosen solution (which previously always was and still is a fountain of interesting and profound discussions) is the structured metadata, collected in an individual metadata schema. This metadata can be used for almost every form of digital content produced at the University of Vienna and all other institutions that are using the system and various other aspects of the long-term preservation of data.

Legal issues are also a very complex subject. The chosen solution addressing such was the involvement in the decision-making progress of a legal consultant specialised in internet law and intellectual property rights. The participation of the labour union of the University was a crucial factor in identifying the appropriate solution for the use of terms to be applied in the Phaidra project.

The integration of Phaidra into other services of the University of Vienna is a major challenge. Especially the connection between Phaidra and further legacy databases and systems as e.g. Fronter are an essential step in the success of the project. Other connections, for example to the projects EOD [1] , WHAV [2] and E-Theses [3] , are established and the progress of integrating objects into Phaidra has started.

But there are also efforts of the university to centralize the numerous different storage systems. Main reasons for this step are centralization, financing and maintenance. As a consequence the number of objects in Phaidra will dramatically increase in the next few months.

2.4 Strength of Phaidra

- Unique features

  The University of Vienna is currently developing an open access policy to motivate researchers to store the intellectual output of the research activities in Phaidra and grant free access to it.

- Access rights

  All persons who have a contract of employment with the University of Vienna as well as all of our students are allowed to upload assets into Phaidra. Guest accounts are included and currently in use. The world is able to view (read-mode) and/or download the assets except in the case where the assets owner restricts access to specific groups of people, individuals or even fully hides the asset. The latter means that only the owner is able to access the object and to modify the metadata. No one is allowed to delete any object.

- Terms of Use

  The Terms of Use stipulate the duties and the rights firstly of the service provider (which is Phaidra) and secondly of the systems users: usage of log files, users commitment to correct conduct (e.g. maintaining awareness of copyright issues), allowance to delete illegal objects, and security issues. Special terms provide regulations in case of the establishment of groups coordinated by a Super-User which holds the maximum amount of rights.

- Licensing

  A person who uploads an object must choose one of six plus one licenses, otherwise they are not able to finalize the upload process. There are six creative commons licenses, the GNU license and one general license available. Finally, users have the option of not choosing any license but keeping all rights reserved.

- Formats

  A document (best practice) informs users about formats that are recommended in order to achieve best permanent digital preservation (see section 3).

- Ease of Use

  Several tutorials and guidelines have been developed to support target groups to properly use the system.

- Training and dissemination work

  In addition, workshops are offered monthly to train people how to use Phaidra. Members of the Phaidra Team are also organizing regular meetings at the faculty level (Phaidra Days), in order to develop the dissemination effort, reach broad acceptance within the university and eventually enhance users' willingness to share learning objects and other materials.

- First-Level and Second-Level Support

  Customer-oriented service is carried out by a Customer Manager. The Help Desk of the University

[1] E-Books on Demand

[2] Western Himalaya Archive Vienna

[3] University of Vienna archive for electronic theses

has been trained for the first-level support. Second-level support is provided through the technical development team. The service website[4] additionally offers extensive information about the system and respective services.

- Updates

  Updates are made once a month.

- Classification of Digital Objects

  Several subject-specific thesauri have already been implemented in order to support indexing.

[2]

3 THE SYSTEM PHAIDRA

As mentioned previously, Phaidra is a digital asset management system with long term preservation aspects[5] . A technical overview will be given in section 4. At this point, we will take a closer look at the system and its limitations.

Every repository stores a huge amount of different data. This data can contain pictures, audios and so on. But there is also the need to store more than one content into one object or sometimes no content at all. So Phaidra differs between three groups of objects:

1. Single-File-Object (one content datastream)
2. Container (multiple content datastreams)
3. Collection (no content - only members of the collection)

With the help of these three-object groups almost every content can be ingested into Phaidra. But these categories are not as precise as desired.

As a consequence, Phaidra differentiates between different object types. These types are Picture, Audio, Video, Document, Resource, Container, Collection, Book, Page, E-Paper and Asset. These eleven different object types are the different content models of Fedora. How these object types are used in Fedora will be described in section 4.1.

To be able to archive and present webpages, Physlets and so on Containers are used. Because of the fact that these type of content has numerous files the Container has been created. Using for these special type a Collection would mean to create numerous objects in the repository that make no sense.

The types ”Book” and ”Page” are special types in the repository which make it possible to produce online books. These books [6] can be viewed via browser (Phaidra Book Viewer) and if OCR data is available there is also the option of setting up a search. But these object types do not accept every file format due to long-term preservation constraints.

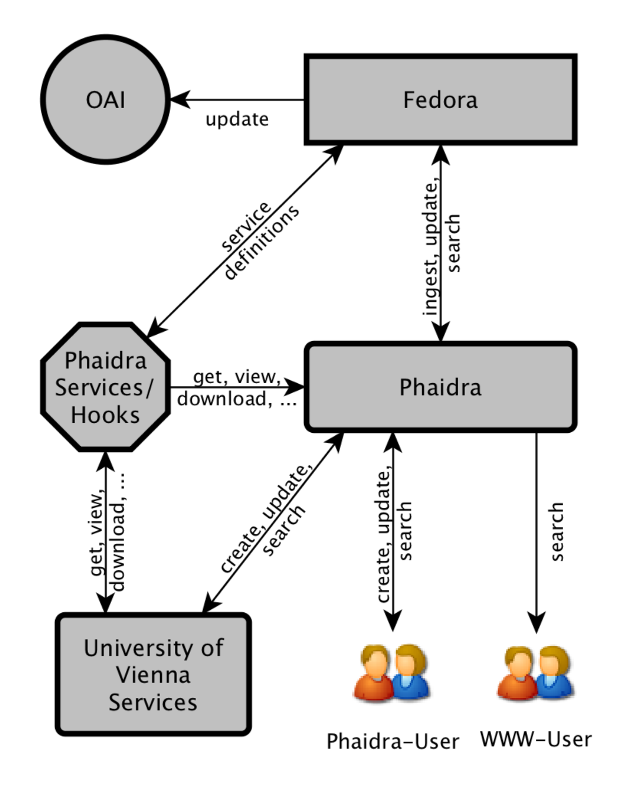


Figure 1. PHAIDRA

Phaidra is designed to migrate the content if it is necessary because in future some types of file formats will be detached by other formats. For this reason, not every content is designed for long-term preservation because of proprietary or inconsistent file formats. Because of this fact, Phaidra only allows a small subset of file formats for the different object types. For example, in the case of the type ”Picture”, the recommended format is TIFF; also allowed are JPEG and JPEG2000; all other formats are not allowed. Thus, if a user wants to upload a GIF image they must choose the Asset type because this special type allows every content type (at the web frontend, this type is called ”Unknown”). So the option exists to upload every content into Phaidra but if the content is declared as an Asset, Phaidra will not take responsibility that it can be accessed in the future.

4 TECHNICAL ASPECTS

PHAIDRA has been implemented expecting 80.000 concurrent users. It can be divided into two parts. Firstly, there is the web frontend that allows members of the University of Vienna to create, update and search objects. Secondly, there is the well-known repository Fedora that is used for storing the objects with their metadata.

In the following sections, the connections between (see Figure 1) these two will be explained.

[4] http://phaidraservice.univie.ac.at/

[5] http://phaidra.univie.ac.at/o:52318

[6] https://phaidra.univie.ac.at/o:19958 - Plinius Historia Naturalis - The oldest book of the University of Vienna

4.1 Fedora

This well-known open source repository is very often used in projects just like Phaidra [5] because it is very reliable and can be easily adapted to special demands [1]. It is implemented in Java and supports features like storing all types of multimedia and their metadata, an API for accessing the repository itself (SOAP and REST), provides RDF search and so on.

Phaidra employs a modified version of Fedora. The modifications to this version took place in different steps. One step was the need of hooks that should check if the submitted metadata is valid. Also a modified search in the repository has been implemented. But all of this was no problem because of the advantages of this repository mentioned before.

Because of the fact that Phaidra uses a modified Fedora the communication with Fedora-Commons/Users is very close. On the one hand because of some kind of bugs and on the other hand because of modifications that may be interesting for the community.

If you integrate an object into the repository, every object will receive a unique identifier. This identifier will never change and so Phaidra uses this identifier as a permanent identifier. For this reason, every object in the repository will be accessible under the same object ID as long as the repository exists.

Fedora uses a Content Model Architecture to differ between the different content that was integrated into the repository and present it to the requester. If you only want to retrieve the content as it is stored, there is no need of using the CMA because Fedora is using its CMA behind the scenes. This architecture is very important for repositories because it is the form of communication of Fedora with other systems.

So Fedora is able to differ between the different content that is stored in the repository. But not every item of content can be returned to the requester as it is saved, because the requester may not interpret the mime type of the content. So the content has to be converted.

Fedora itself is not able to convert content, but with the help of the CMA and the possibility of defining services, this problem has been solved.

These service definitions (see Figure 2) represent the ways Fedora communicates with other services, for example, converting the content from format X to Y. Every content model in Fedora has service definitions and tasks to do different jobs that can be defined by the administrator of the repository.

A service definition defines services and different operations. To carry out the requested job, you must also define service tasks because the service definition will not know how to interpret the service outside of Fedora. The defined service deployments can be linked to the different service definitions and so Fedora knows what to do if a certain operation is requested.

There are different ways of searching in Phaidra. First, you can search in the index with the help of SPARQL. Before Fedora 3.3, this method was sometimes very inefficient because it was very slow. Now, searching in the index is about ten times faster. For example, the OAI-PMH provider can be used to search for new created or updated objects.

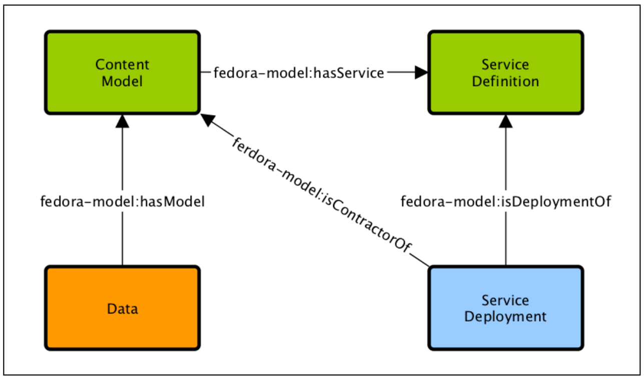


Figure 2. Fedora CMA Relationships

In addition, a full-text search is also available. It is called GSearch and with the help of this, you are able to search in defined fields of the index. To perform the search, Lucene is used because it is quite fast and common. So if you configure Fedora to extract the fulltext of the PDFs during integration, you will then be able to search through all documents and receive a result of the matching hits.

Actually about 60.000 objects are in the repository available. These objects are saved on the main SAN of the University of Vienna and use about 2 Terabyte of space. Most of the objects are Picture (about 20.000) and Page (about 30.000) objects.

4.2 Phaidra

4.2.1 Phaidra Core

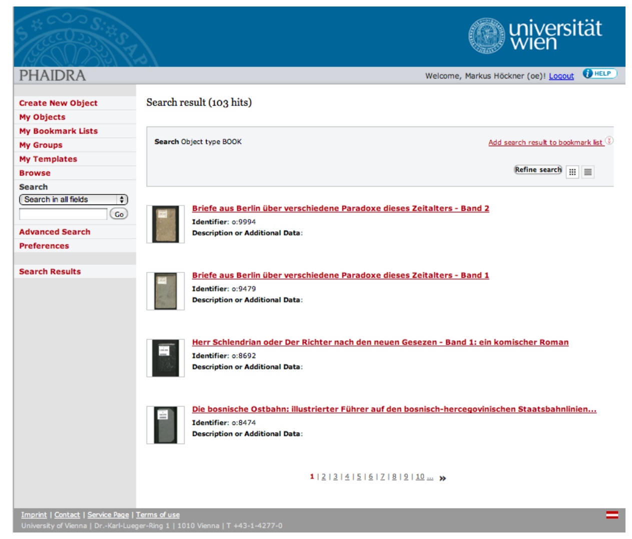
There are several interfaces for Fedora available just like Muradora and so on. But the University of Vienna decided to create an own interface. Main points for this decision were

- adapting the interface to the internal structure of the university,
- adapting the interface to existing ones,
- designing an ”own” interface.

Because of the fact that most of the web applications at the University of Vienna are written in Perl the programming language of Phaidra Core is also Perl. So until 2013 the Core will be developed in Perl. But the Phaidra Team traces the development of the other interfaces with great interest.

To assure scalability, extensibility, reusability, flexibility and reliability the web application framework Catalyst [6] is used.

Catalyst is a widespread framework for Perl and has a very strong community. Because of the fact that a lot



Search result (103 hits)

Search Object type BOOK

Briefe aus Berlin über verschiedene Paradoxe dieses Zeitalters - Band 2
Identifier: o:9994
Description or Additional Data:

Briefe aus Berlin über verschiedene Paradoxe dieses Zeitalters - Band 1
Identifier: o:9479
Description or Additional Data:

Herr Schlendrian oder Der Richter nach den neuen Gesezen - Band 1: ein komischer Roman
Identifier: o:8692
Description or Additional Data:

Die bosnische Ostbahn: illustrierter Führer auf den bosnisch-hercegovinischen Staatsbahnlinien...
Identifier: o:8474
Description or Additional Data:

University of Vienna | Dr.-Karl-Lueger-Ring 1 | 1010 Vienna | T +43-1-4277-0

Figure 3. Phaidra Webfrontend

of web applications work with Catalyst, there are a lot of plugins available for it. For this reason, the developers do not have to implement the web application from the scratch. They can then concentrate on other issues that may arise.

For integrating objects, updating the metadata or searching in the repository SOAP is used. Since Fedora 3, there is also the possibility of using REST, but when the project started this option was not available. So Phaidra must be connectable to a specific API-A or API-M method of Fedora and put the data into Fedora or retrieve it.

For character encoding, Phaidra uses UTF-8. Because Phaidra is able to handle almost every language and their lettering in the metadata and webfrontend. So there is no need for a coding or decoding location that would otherwise cost great performance.

The webfrontend (see Figure 3) is fully localized in German and English and soon there will be an Italian version available. When the web frontend was developed, the usability of the application was very important. So the application was designed as simple as possible and clearly structured. To fulfill these demands, new web technologies like AJAX are used.

But there is also the need for batch or automatic uploads. To carry these out, a special Phaidra API was developed by the Phaidra team. If a user wants to upload objects into Phaidra they do not have to use web frontend. The creator of the object is able to use this API to create and update objects in Phaidra. Also, a search method is available. And now this API is available for Perl and Java.

4.2.2 Phaidra Services/Hooks

The Services and Hooks of Phaidra are very important parts of the web application because they have to perform much work.

To be able to convert pictures, to view them in the browser, to download a PDF document or to play a video, other services are required as Fedora is not able to do this work. So not only the web frontend and the ingest of objects are main developing points, also the services are important. They are responsible for presenting the content in the appropriate way.

For example, a picture is uploaded into Phaidra and you would like to view it in your browser. There might be the problem that the uploaded content is a TIFF image. So without a browser plugin you are not able to view it. To prevent that every user is able to view this picture in the browser it has to be converted in a format that every browser can interpret, for example, JPEG.

So the first step is that Fedora recognizes that you need a JPEG image. For this, the CMA of Fedora is used. Fedora recognizes the object as a picture due to the CMA. With the help of the service definition, the system connects the object to the appropriate Phaidra Service. The system calls this service to convert the picture into JPEG format.

In Fedora, almost every kind of data can be stored. So there is no problem with saving metadata in an XML and adding this XML as a datastream to the object. The problem is that if you have a metadata schema, you must also define a structure for the XML as well as special vocabulary. To ensure that only valid XML datastreams are added to objects in Fedora, Phaidra has so-called Hooks. These Hooks are responsible for checking if the metadata is valid and the object has all of the required datastreams and service definitions. For this reason, these hooks are possibly the most relevant parts of Phaidra because they guarantee reliability and security.

5 UNIVERSITY OF VIENNA METADATA

Metadata is structured data about other data and fulfills a variety of tasks [3].

- identify objects worldwide;
- describing objects (e.g., author, creator, title, description);
- support of information retrieval and identification;
- describing the historical audit trail of an object and its provenience;
- grouping objects into collections;
- rights information, licenses and access permissions;
- technical information about the content;
- easier interchange of data between autonomous repositories;
- versioning of an object;

Because of this metadata, the objects are somehow self-documenting and prepared for long-term preservation. But most of the metadata must be created by humans. As a consequence, metadata costs a lot of money and time to preserve every object in the best way.

The metadata schema of the University of Vienna is a modified LOM schema. LOM is a standard by the IEEE [4] that is well known for describing and documenting learning objects in a repository. The standard describes the basic structure of the metadata, datatypes, list values and vocabularies. The need of such a standard is significant because of interoperability and long-term preservation.

The first version of Phaidra contained the LOM schema with some specific adaptions just like extending the vocabularies. After a certain time, faculties of the University of Vienna got in contact with the Phaidra team and the metadata working group. The reason for said contact was that they also had data to store but they needed specific metadata. As a consequence, the LOM schema had to be extended again. So the process of analyzing and adapting started again and still continues.

In the last two years, two major adaptations were made on the metadata schema of the University of Vienna. First, there was the need to store primary data. Especially the Institute History of Arts at the university asked for this new section. Thus, cooperation between Phaidra and this institute was initiated.

Phaidra also stores digital books and so book-specific metadata (e.g., publisher, publishing date) had to be included into the schema. In cooperation with the Library of the University of Vienna, these new metadata tags have been added and included.

To be able to offer Dublin Core metadata, the University of Vienna metadata schema has eight mandatory fields. So the user of Phaidra does not have to add Dublin Core metadata manually because it is extracted automatically from the metadata schema of the University of Vienna. The Dublin Core will be saved into the object as a datastream so that it is easily accessible.

6 FORECAST

The project development of Phaidra has been extended by the University of Vienna for the next three years because of the need of such a repository and its applications. The main step in the project will be to set up connections to other systems at the university and to migrate data into the repository.

Besides the improvement of the existing streaming services there is also a need for collaborative functions for the SuperUser and for a new ImageViewer. This new application will allow users to view images greater than one hundred megabytes via the WWW. The basic elements for this new application exist (Phaidra Imagemanipulator) but there is still much work to do.

Phaidra also participates at several Europe-wide projects. The two best known are TEMPUS and OPENAIRE.

TEMPUS (Trans-European Mobility Scheme for University Studies) is a European Union programme. The aim of the project is the modernization of higher education in countries surrounding the EU. Through this project the University of Vienna stays in touch with several Western Balkan universities.

OPENAIRE (Open Access Infrastructure for Research in Europe) deals with the problem of open access at a European level. The project currently has 38 partners from 27 European countries.

One of the oldest universities of the world, the University of Padua, has contacted the University of Vienna due to the need of a repository. They from University of Padua took a closer look at Phaidra and decided to join the project. So Phaidra has been installed there at the beginning of May 2010.

Also, national universities and institutions have contacted the Phaidra group. Until now, three instances of Phaidra are planned, one for the University of Applied Arts Vienna, one for the University of Music and Performing Arts Graz and another for the Austrian Science Board. Phaidra is also in contact with two more Austrian universities.

Since the end of 2009, Phaidra is also a part of the project Europeana. The aim of this project is to make digital content efficient and quickly accessible at a European level. The first objects from Phaidra have been successfully transferred to Europeana. To enable this, the OAI-PMH provider of Fedora is used.

So the development of Phaidra is not finished yet. With the help of other resources, the project will grow in the next few years. Also, sharing of know-how with other universities and projects will help in the development Phaidra.

7 REFERENCES

[1] Bainbridge D. AND Witten I.H. ”A fedora librarian interface”, *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* , Pittsburgh, U.S.A, 2008.

[2] Budroni P. ”Manifest zur Bildung einer Matrix”, *Mitteilungen der Vereinigung sterreichischer Bibliothekarinnen und Bibliothekare 63 (2010) Nr. 1/2* , Bregenz, Austria, 2010

[3] Gladney, H.M. *Preserving Digital Information*. Springer, Berlin-Heidelberg, 2007.

[4] Institute of Electrical and Electronics Engineers, Inc. *Draft Standard for Learning Object Metadata*. New York, 2002.

[5] Kumar A., Saigal R., Chavez R. AND Schwertner N. ”Architecting an extensible digital repository”, *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* , Tuscon, U.S.A, 2004.

[6] Rockway, J. *Catalyst - Accelerating Perl Web Application Development*. Packt Publishing, Birmingham - Mumbai, 2007.

[7] Waugh A., Wilkinson R., Hills B. AND Dell’oro J. ”Preserving digital information forever”, *Proceedings of the fifth ACM conference on Digital libraries* , San Antonio, U.S.A, 2000.