Integrating Planets and Fedora Commons

A Case Study of Integration of Planets Characterisation Services with the Digital Object Management System of The State and University Library, Denmark

planets

Executive Summary

The State and University Library, Denmark houses, among other national collections, the national media and newspaper collections. Its digital collection is growing and the Library is implementing a new Digital Object Management System (DOMS) to replace over forty legacy repositories in which it is currently held. This new system has Fedora Commons at its core.

While the Library wishes to characterise the files in the digital collection before moving them into DOMS, this functionality is not provided by Fedora. Planets was chosen to provide characterisation services and they now have a proof-of-concept system which has integrated Planets into its workflow for loading the digital collection into DOMS. Planets characterisation results are stored with the digital objects in Fedora.

Work being done on DOMS has prompted the Library to review its digital preservation policy and strategy, and once this is completed they will take a decision on how much a part Planets will play in their ongoing digital preservation solution.

Authors
Lynne Chivers, The British Library
Asger Askov-Blekinge, State and University Library, Denmark
Bjarne Andersen, State and University Library, Denmark

This case study is part of a series of case studies on the application of Planets in major European libraries and archives.

They are all available via the Planets website **www.planets-project.eu**

The State and University Library, Denmark

STATSBIBLIOTEKET

As a national library of Denmark, the State and University Library contributes to collecting and providing access to Danish cultural heritage and preserving this for future generations.[1]

The 20th century is the era of a fast growing media industry. From the invention of sound recording and film at the end of the 19th century to today's internet, the development has been immense. Introduction of broadcast in the first half of the last century and the explosion of the music industry in the second, together with the continuous appearance of new consumer media formats (vinyl records, cassette tapes, video tapes, compact discs and DVDs) have had an enormous impact on society. The media content, ranging from small local events to global breaking news and covering all kinds of cultural manifestations, constitutes our audiovisual memory.

A great part of this section of Danish cultural heritage is housed within the State and University Library. The holdings cover among other things some of the oldest sound recordings in the world in terms of unique wax cylinders from the 1890s, a nearly complete collection of all Danish gramophone records, radio and television programmes and advertising films from the 1950 and on. Parts of the audiovisual collections are of international importance, for example private recordings of international opera singers and possibly the earliest recording of the Swedish national anthem.

From the audio/visual collection facilities showing some reel-tapes with older radio-recordings, State and University Library.

State and University Library, Aarhus, Denmark. Photo AU-foto

Part of the national CD collection, © Thomas Søndergaard

Important sources of culture and history are to be found in these collections. The archive is a valuable and well used source for research and education, and several research projects have emerged on the basis of the collections. Also, there is a strong demand from museums to include audiovisual media to enrich their exhibitions.

1 http://en.statsbiblioteket.dk/

«The Ruben collection of old wax cylinders[2] at the State and University Library offers a fascinatingly rich impression of the vibrant cosmopolitan cultural and musical life of Copenhagen in the 1890s.

The repertoire preserved in the Ruben collection is broad indeed highlighting the cultural appropriation of not just German, French and Italian music but also for instance American popular song. A wonderfully mixed bag of music ranging from the very first recordings of the Danish royal and national anthems to American minstrel song, and from Verdi opera to the latest vaudeville hit. Not to mention a reading of a Hans Christian Andersen story, another first.

In short, this heritage collection constitutes an invaluable and globally unique historical documentation, which gives us a wonderfully varied soundtrack to a hitherto totally silent era in Danish cultural and musical history.»

Steen Kaargaard Nielsen, Associate Professor, Ph.D.
Department of Musicology, Aarhus University

The Library has a legal responsibility to preserve these media, and there is only one way to do this: digitisation. This is because some original media are very fragile and in many cases are no longer playable. In other cases, the Library has only ever held a digital version. Long term digital preservation is therefore vital for the survival of these cultural treasures.

«Long term digital preservation is a necessary condition for the survival of these cultural treasures.»

Eva Fønss-Jørgensen,
Head of National Collections,
State and University Library.

2 See http://www.youtube.com/watch?v=edqr0SlScZc for a video film about the Ruben Collection.

The Digital Collection

The State and University Library houses the national media and newspaper collections. There are 180,000 physical items of audio and video material, held on analogue carriers such as VHS tape, DAT tape, reel tape and wax cylinders, and around 76 million pages of newspapers. The Library is currently running or planning massive digitisation projects on these collections. For instance they are in the planning and fundraising phase of a project to digitise 12 million newspaper pages. In addition, the Library has been recording and on-the-fly digitising Danish radio and television broadcasts from analogue, cable, satellite and internet since 2005.

For example, in its digital collection the Library currently has:

- Two million pages of digitised paper materials.
- 1.2 million hours of Danish Radio/TV in wav, mpeg-2 and mpeg-4 formats (550 TB).
- 450,000 tracks from Danish CD's, in wav format (18 TB).
- 45,000 advertising films from commercial television and cinema.

The digital collection is expected to increase in volume by 240 TB in 2010.

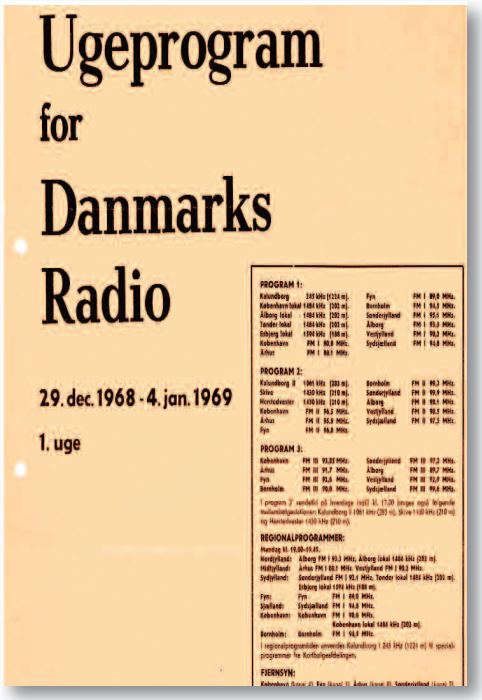
Digitisation of wax cylinders at the State and University Library. The originals are very fragile – so it is very important to keep the digital versions safe. In many cases the originals are not playable any more.

«Many of the digital collections are unique, so securing these materials in the long term is essential to the overall task of preserving Danish cultural heritage.»

Bjarne Andersen, Head of Digital Resources, State and University Library, Denmark

Kiøbenhavns
Kongelig allernaadigst privilegerede
Adresse-Contoirs
Efterretninger.
Vi og fyrretyvende Aargang
1807.
Kiøbenhavn.

Front cover of newspaper from 1807, State and University Library



Ugeprogram
for
Danmarks
Radio
29. dec. 1968 - 4. jan. 1969
1. uge

Front cover of a Danish TV-guide, State and University Library

DOMS, Fedora and the Need for Characterisation

The digital collection at the Library is held on over forty legacy "repository" systems. To make the collection easier to maintain and use, they wish to migrate these old systems to a new, more generic system, capable of holding any kind of digital content. To achieve this they decided to build a new Digital Object Management System (DOMS) for long term preservation, with Fedora Commons at its core[3]. The Library is a contributor to Fedora and they have developed extra functionality to be used in DOMS. See Figure 1 for an outline of the repository architecture.

The existing digital collection is held in a wide variety of formats. Before ingesting to the new DOMS, the Library would like to characterise and validate the files to ensure that only valid files are loaded (ie those which conform to the specifications of their file formats). In addition, they may wish to normalise the files to a small, selected set of formats chosen as eligible preservation formats permitted to be stored in DOMS for long-term preservation.

The vast number of files concerned means that automating the characterisation and validation of them is essential. However Fedora does not provide the full characterisation and migration functionality required by the Library. They needed to look elsewhere for a solution.

What is Fedora?

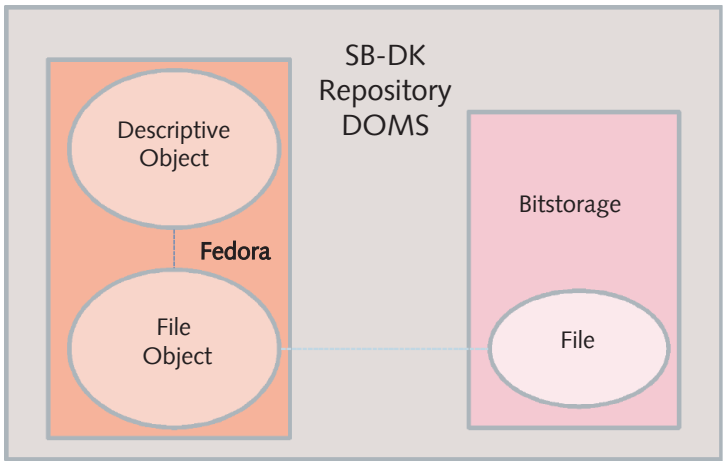
The Flexible Extensible Digital Object Repository Architecture (Fedora) is a conceptual framework that provides the basis for software systems that can manage digital information. The Fedora Repository is very flexible and is capable of serving as a digital content repository for a wide variety of applications, such as digital libraries and archives, institutional repositories and content management systems. It is able to store any type of digital content, such as documents, images, video, plus metadata about the content items in any format. In addition, the relationships between the content items can be stored. It is possible to store just the metadata and relationships for content which is held by another system or organisation. Content items can either be stored locally in the repository, or stored externally and just referenced by the Fedora digital object.

The Fedora Repository is a product of Fedora Commons, a non-profit organisation, and is provided as free, open-source software. While it is capable of operating as a standalone content server, it is really designed for use with other software. In most cases it will be only part of a complete content solution incorporating other components such as ingest applications, search engines, workflow management and security.

For further information about Fedora see the Fedora Tutorial, Introduction to Fedora on the Fedora Commons website.

The Descriptive Object contains:
- Descriptive metadata
- Reference to relevant File Object

The File Object contains:
- Technical metadata, e.g. file format
- Reference to relevant permanent URL of File in Bitstorage



The File is:
- The digital collection item to be preserved, e.g. audio file or scanned image file

Figure 1 Overview of the DOMS Repository based on Fedora Commons at The State and University Library (SB-DK).

3 http://fedora-commons.org

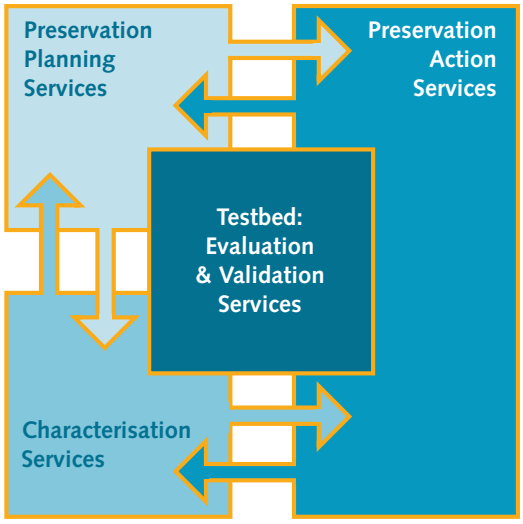
The Solution

The State and University Library is a partner in Planets, a European joint-venture project for research and development in the field of digital preservation, which has produced a framework and set of practical tools and services to enable institutions to manage and access digital collections for the long-term[4].

Included in the Planets suite of services are tools to characterise files, which can identify and validate many commonly-used file formats. The characterisation services draw on the Planets Core Registry which contains technical information about file formats and their properties, as well as the preservation actions that might apply to them.

The Planets Interoperability Framework unites the full range of services provided by Planets, and can be integrated with external services such as common archiving and library systems. This is illustrated in Figure 2. Integration is achieved via an adapter to call on specific services contained within the Planets Framework from the external system. Once an adapter is incorporated within the external system, tools in Planets can be run from any platform. There are two parts to the adapter: one provided by Planets (labelled Planets Connector in Figure 2) and the other specific to the particular external system (labelled External Connector in Figure 2).

What is Planets?



Planets makes it possible to:

- Define preservation policies and goals
- Assess the preservation needs of an organisation, collection and users
- Identify areas where preservation of collections does not meet policy requirements
- Build, evaluate and execute plans to address any problem areas
- Analyse and verify the results
- Document the decisions made and actions taken

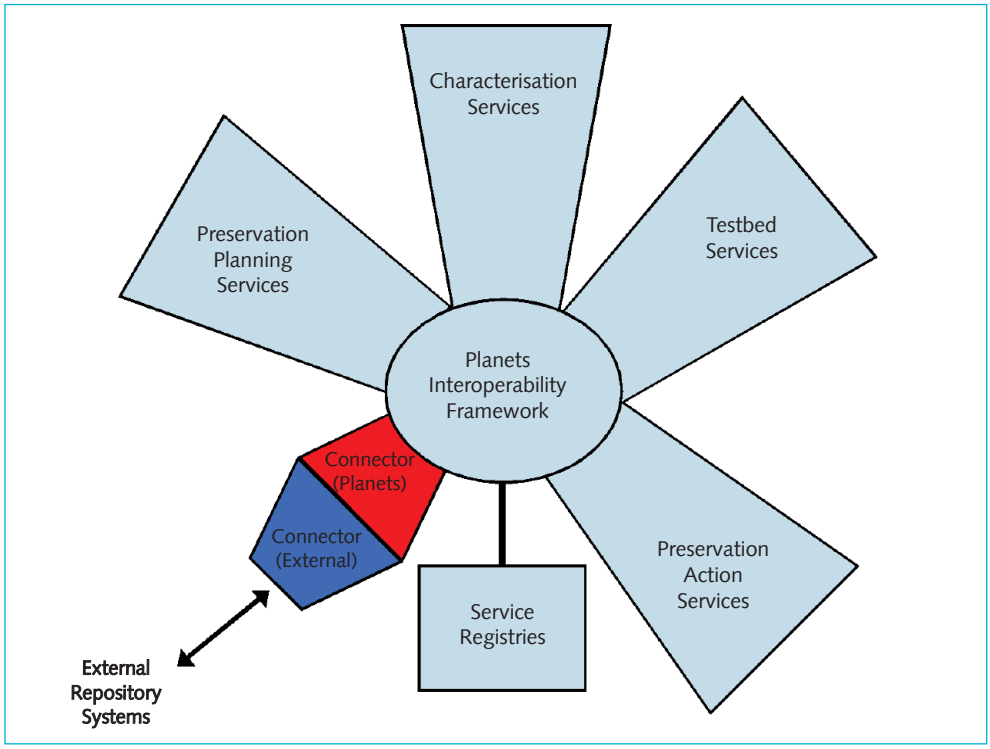


Figure 2 Overview of Planets Services and Interoperability Framework which unites them, showing the two parts of the adapter to external services.

4 For more detailed information about Planets tools and services see http://www.planets-project.eu/docs/comms/PLANETS_BROCHURE.pdf and http://www.planets-project.eu/docs/comms/PLANETS_PRODUCT_SPECIFICATION.pdf

The Library decided to set up a proof-of-concept system which would integrate Planets characterisation services into an automated workflow for digital collection items to be loaded ("ingested") into DOMS. They used a test version of DOMS, installed a local Planets server and built a Planets/Fedora Connector. They built workflow software to manage the process of loading a digital object to DOMS, including the characterisation of the file and evaluation of the characterisation result. Evaluation checks whether the file's format has been identified, whether the format has been validated and whether the format is one which has been deemed to be an acceptable format in DOMS. (A set of acceptable formats is derived from the preservation strategy and is accessible by the Evaluator.)

The work to build the system was completed over a period of about five months, with around six weeks of coding effort. The workflow software is a webservice, written in Java and the Connector is a Java library.

The automated process of ingesting a digital file object to the DOMS repository is illustrated in Figures 3 to 6.

In Figure 3, the File and its Descriptive Object are passed to the workflow software (the Workflow Engine) and the ingest process begins (1a). The File is stored in temporary storage within DOMS (1b) and is allocated a permanent URL. This URL is passed back to the Workflow Engine (1c). The Descriptive Object is stored in Fedora and the permanent URL is stored in the related File Object in Fedora (1d).

«Planets automates the process of identifying the characteristics of the digital materials we wish to preserve.»

Adrian Brown, The Parliamentary Archives, UK.

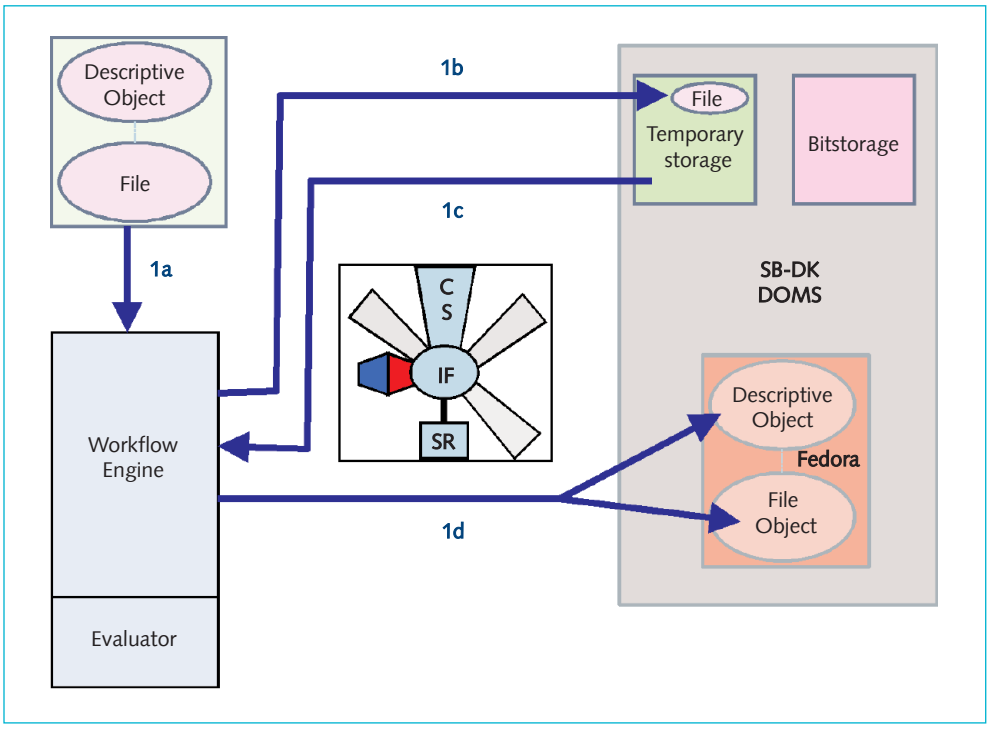
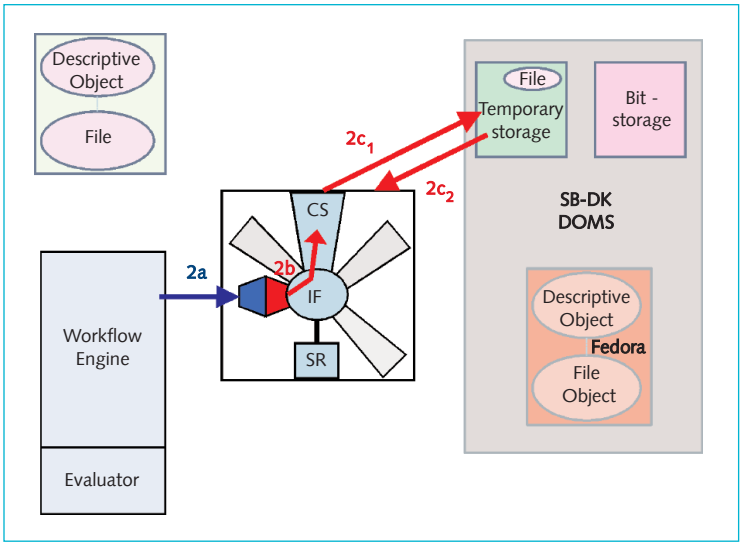
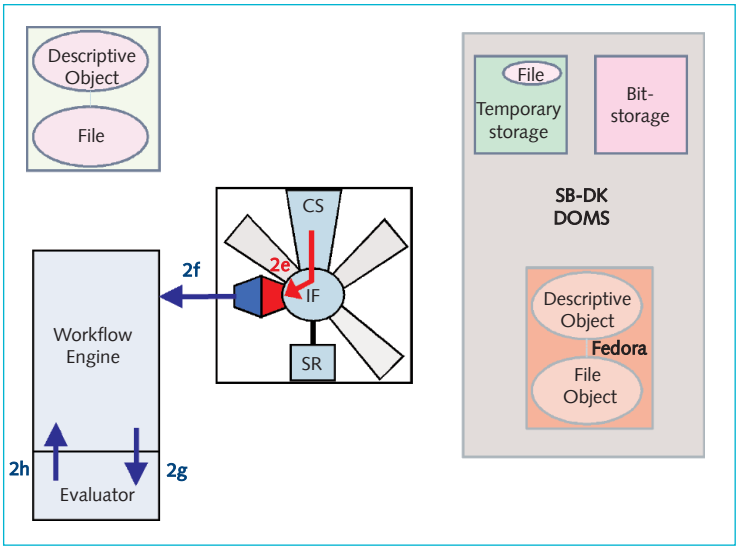


Figure 3 Initial loading of digital object to DOMS.



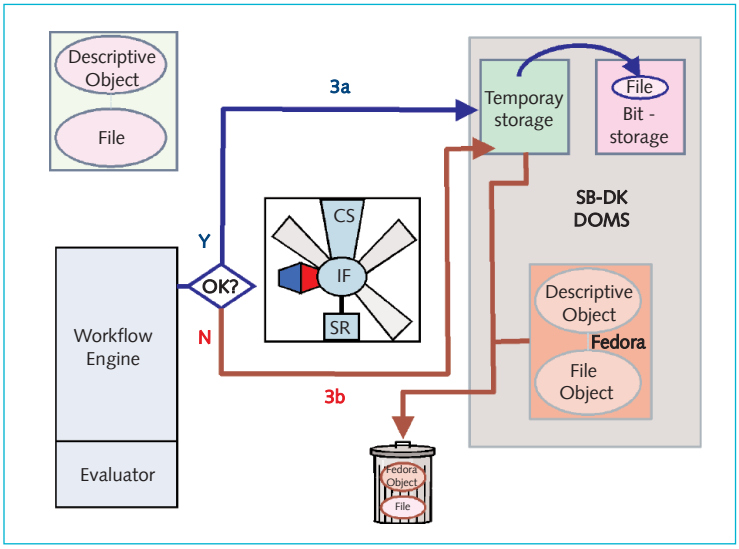
Now that the object is stored in DOMS, the Workflow Engine initiates characterisation of the File via Planets, and this is shown in Figure 4. The Workflow Engine requests characterisation via the adapter (2a). Planets calls the required characterisation services (2b) which fetch the File from Temporary Storage (2c1 and 2c2), and characterisation takes place. Examples of characterisation services are Droid and Jhove[5].

Figure 4 Characterisation of the File.



The evaluation of the characterisation result is shown in Figure 5. The result is returned via the adapter (2e) to the Workflow Engine (2f). The Workflow Engine requests evaluation of the characterisation result (2g) and the outcome is returned to the Workflow Engine (2h).

Figure 5 Evaluation of characterisation result.



The final step of the ingest process is shown in Figure 6. If the evaluation outcome is that the File is valid and may be ingested to DOMS (OK=Y), then the File is moved from Temporary Storage to permanent Bitstorage in DOMS (3a). Otherwise (OK=N), the File is removed from Temporary Storage and the entire object is removed from Fedora (3b).

Figure 6 Complete the ingest process or rollback, depending on outcome of evaluation.

5 http://sourceforge.net/projects/droid
http://sourceforge.net/apps/mediawiki/droid
http://hul.harvard.edu/jhove

The Outcome

The Library now has a proof-of-concept system to ingest the digital collection to DOMS, with automated characterisation provided by Planets characterisation services. The system will evolve to become the live system used in the Library. The system is Planets-ready, although not Planets-dependent.

The Library has issued the code they have developed as a Fedora plug-in for Planets integration. This is open source and has been shared with the Fedora Community[6].

Future Developments Using Planets

The work being done on DOMS has prompted the Library to review its digital preservation policy and strategy. Once these have been agreed, it will consider how much a part Planets will play in its eventual digital preservation solution. One possible scenario is shown in Figures 7 to 11, which illustrate an automated normalisation process for digital objects being stored in the digital repository. Normalisation ensures that all files entering the repository conform to a pre-defined set of formats which have been deemed by the policy as being acceptable as a preservation format at the time of ingest.

The first step in this process is shown in Figure 7. The File and its Descriptive Object are held in a pre-ingest storage area and the process is initiated (1a). The Workflow Engine requests characterisation via the adapter (1b). Planets calls characterisation services (1c). Planets uses its Service Registry (SR) to identify appropriate characterisation services either within Planets or across the web. The File is fetched from Temporary Storage (1d) and characterisation takes place. The characterisation result is then stored in the Descriptive Object (1e).

Review of Preservation Policy and Strategy

The State and University Library is currently working on defining a strategy for preservation of its digital assets. The work on DOMS has resulted in the need for the strategy, in order to steer the development of DOMS in the right direction. It is necessary to define and place responsibilities as well as encourage ownership of the preservation process right from creation of the digital object.

In 2002, the Library was part of a national study which focused on preservation of digital as well as physical library and museum archives. The study listed a number of options for digital preservation but the Library did not take a stand on which approach to take. However, inspired by the Library's participation in the Planets project, the IT section Digital Resources, which covers functions such as digitisation, software development and research in long term digital preservation, has been working on a strategy since autumn 2009.

A strategy group was established, comprising a cross-disciplinary team of curators, software developers, librarians, technical experts, etc, and a workshop was held to ensure all participants had the same foundation for the discussion. A reading list of documents was produced, including for example the LIFE model[7] and the British Library Digital Preservation Strategy[8]. This workshop was followed up by a brainstorming session.

The next step is to organise the input from the brainstorm, research further, and produce a first draft of the strategy for internal review by the group. The final version of the strategy will be presented to the Library's management. The goal is to have the final strategy in place by the end of 2010. Concurrent with formulating a strategy, work will be carried out to define the Library's overall policy on digital preservation.

6 http://sourceforge.net/projects/planetsfedora/
7 The LIFE project has developed a model for estimating the preservation costs of a digital object's full lifecycle. http://www.life.ac.uk/
8 http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresstrat.pdf

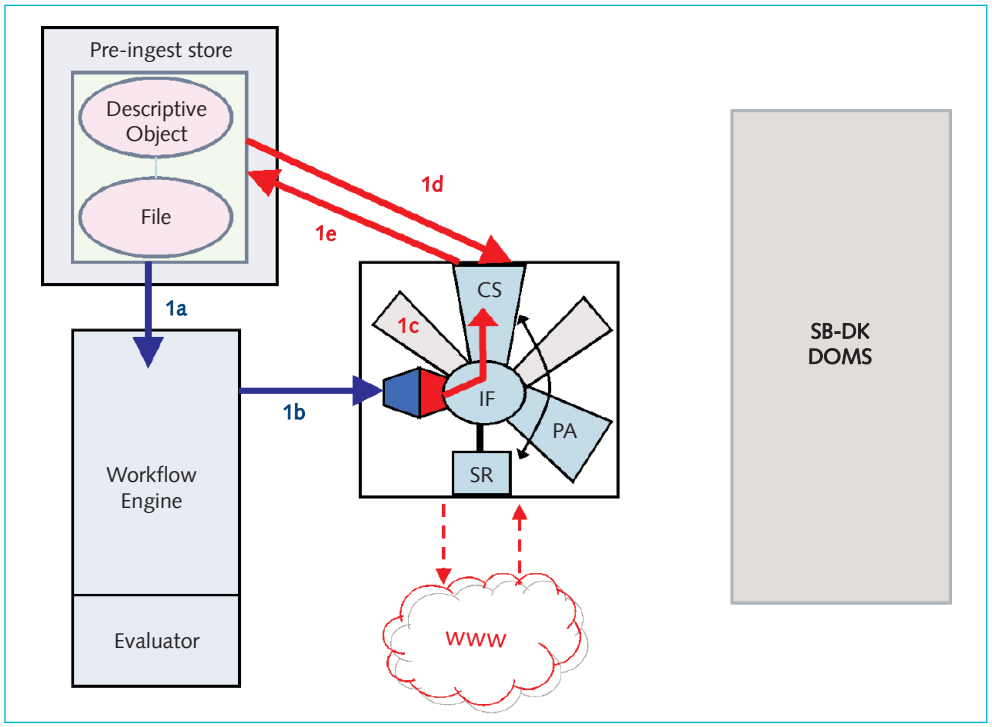


Figure 7 Characterise File via Planets prior to ingest into DOMS.

The next step is to evaluate the result of the characterisation, and this is shown in Figure 8. The characterisation result is passed via adapter (2a) to the Workflow Engine (2b). The Workflow Engine requests evaluation of the characterisation result (2g) and the outcome is returned to the Workflow Engine (2h). (Note that this step equates to that shown in Figure 5.)

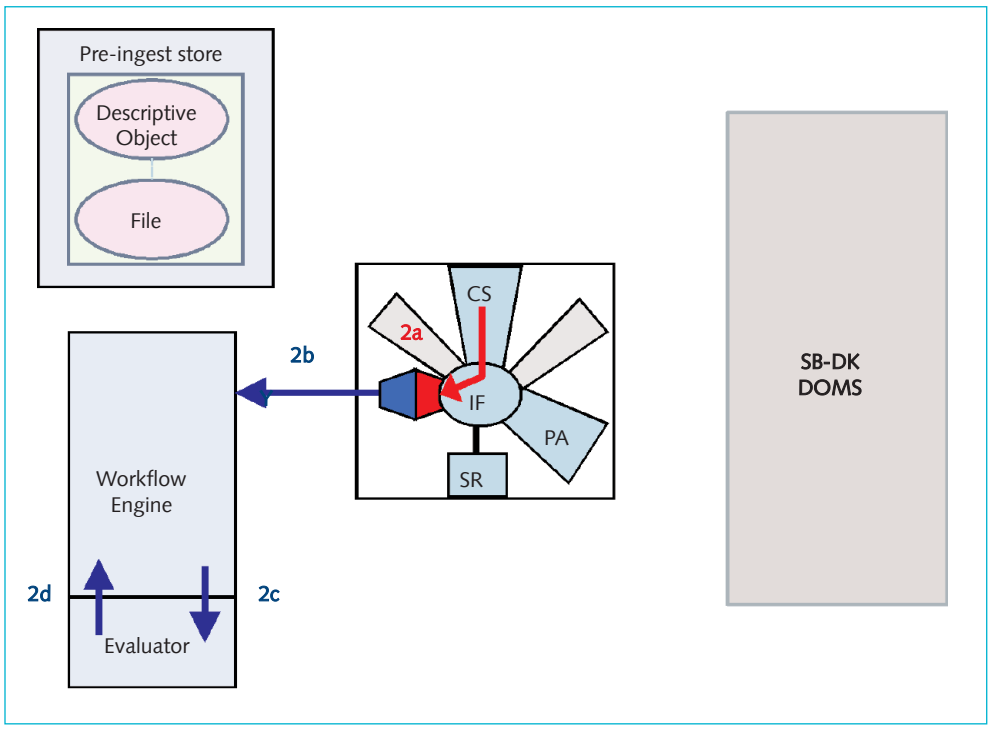


Figure 8 Evaluation of characterisation result.

If the outcome of the evaluation is that the File is valid and is acceptable as a preservation format in DOMS, then the File and its Descriptive Object may be stored in DOMS. This is illustrated in Figure 9. The Workflow Engine fetches the File and its Descriptive Object, which by now contains the characterisation result (3a). They are then stored in DOMS and the ingest process for this File is now complete.

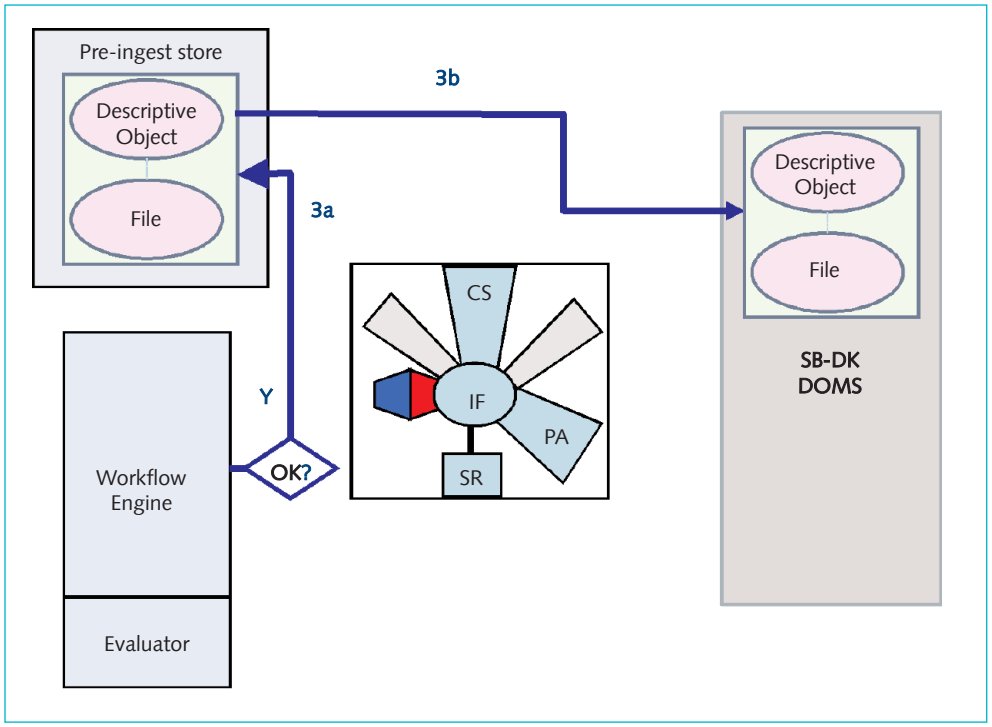


Figure 9 The outcome of evaluation is that no normalisation is required so the File and its Descriptive Object may be stored in DOMS.

If, however, the outcome of the evaluation is that the File's format is not acceptable as a preservation format in DOMS, then normalisation is required. For this scenario the remainder of the process is illustrated in Figures 10 and 11.

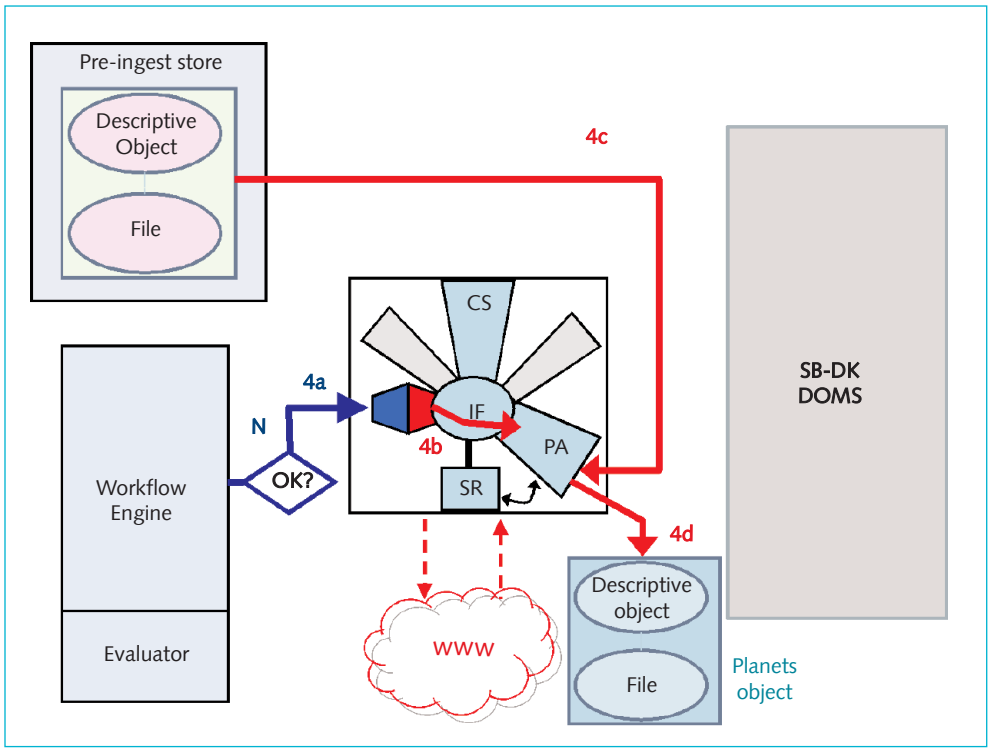


Figure 10 Migration of File to normalise its format ready for ingest to DOMS.

The normalisation of the File is shown in Figure 10. The Workflow Engine requests a migration of the File to a normalised format via the adapter (4a). The policy for this is machine readable so the Workflow Engine knows that, for example, MP3 files must be normalised as BWF format. Planets performs migration via Preservation Action Services (4b)[9]. The File and Descriptive Object is fetched from the Pre-ingest store (4c), and an appropriate migration tool is used to create a new version of it in the specified normalised format (4d). The new version is a Planets Object and contains a record of the migration which took place. Planets uses the Service Registry (SR) to identify appropriate migration services, which may either be in Planets or remote across the web.

The final step of the process where normalisation was required is shown in Figure 11. Having completed the migration of the File to the new, normalised format, Planets returns the reference for the new Planets Object to the Workflow Engine (4e). The Workflow Engine moves the Planets Object to DOMS (4f and 4g) and stores the original in DOMS (4h and 4i).

«Colleagues at Cornell University and DuraSpace[10] are studying Planets very intently to inform the National Science Foundation Data Conservancy projects in addition to our work in Fedora.»

Daniel Davis, Cornell University and DuraSpace Affiliate.

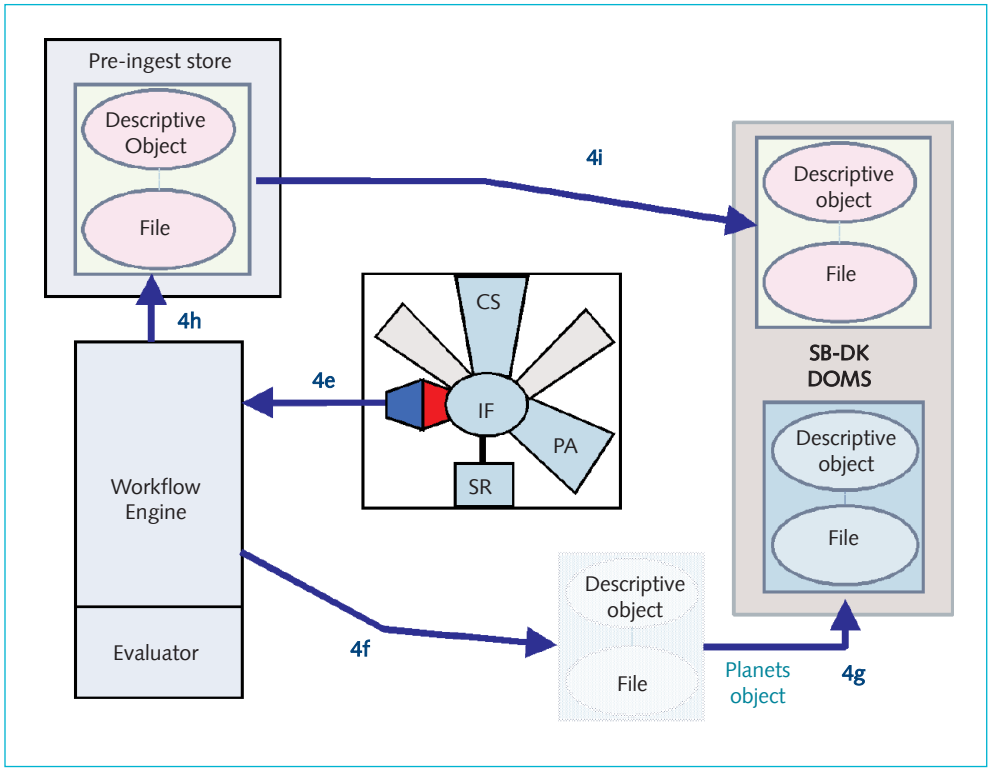


Figure 11 The reference to the new Planets Object containing normalised File is returned to the Workflow Engine and the original and normalised objects are stored in DOMS.

9 Planets Preservation Action Services include services for migration (converting files from one format to another) and emulation (enabling files to be accessed via old operating environments).

10 This work is being undertaken by Chris Wilper, Aaron Birkland and Daniel Davis who are committers on Fedora and Thornton Staples who co-started the Fedora project. Committers contribute to the Fedora open source software.

Conclusions

The State and University Library holds a digital collection of great cultural value that must be preserved. Its Digital Object Management System, based on Fedora Commons and being developed by the Library, will be used to store this collection. Planets can help them to preserve the items contained within it.

They have demonstrated that Planets services can easily be integrated with existing repository systems, having done so with their Fedora-based digital repository. Planets characterisation services have been used to evaluate digital objects to be ingested into the repository.

Code developed by the Library, the workflow/evaluation and the Planets/Fedora Connector are open source and have been shared with the Fedora Community[11].

The Library is engaged in a review of its preservation policy and strategy and Planets is expected to provide part of their ongoing digital preservation services. It is a founding member of the newly-established Open Planets Foundation.

«We would never be able to solve a global problem like digital preservation on our own. This is why cooperations like Planets are extremely useful.»

Bjarne Andersen, Head of Digital Resources, State and University Library, Denmark

11 http://sourceforge.net/projects/planetsfedora/

References and further reading

Fedora Tutorial, Introduction to Fedora, The Fedora Development Team, 2008.
http://www.fedora-commons.org/confluence/download/attachments/4718930/tutorial1.pdf?version=1&modificationDate=1218459761506

Digital Preservation – the Planets Way; summaries for technical/developer staff, Raymond van Diessen (IBM), Laura Molloy and Andrew McHugh (both HATII, University of Glasgow), 2010
http://www.planets-project.eu/training-materials/IntroductiontoDigitalPreservation-TechnicalSummary-Final.pdf

An overview of Planets is given in the Planets brochure at
http://www.planets-project.eu/docs/comms/PLANETS_BROCHURE.pdf

A more in depth look at Planets tools and services may be found at
http://www.planets-project.eu/docs/comms/PLANETS_PRODUCT_SPECIFICATION.pdf

Acknowledgements

The authors would like to thank all those who have contributed to this case study.

Front cover images:
State and University Library building in Aarhus, Denmark. Photo Thomas Søndergaard and AU-foto.
Example frame from television broadcast collection of the State and University Library.

Bjarne Andersen, page 19 , Photo © Thomas Søndergaard

Planets (Preservation and Long-term Access through NETworked Services) is a four-year, €15 million project, co-funded by the European Commission under the Information Society Technologies (IST) priority of the 6th framework Programme (IST-033789).

The project has developed a suite of tools and services to support preservation of digital content for the long-term. Planets tools make it possible to define digital preservation goals and policies; understand the characteristics of a collection; build, evaluate and execute preservation plans, convert objects into up-to-date and accessible formats and run software on legacy operating systems. It offers an automated solution to support informed decision-making and justify actions taken.

Planets is coordinated by the British Library and has been delivered by a Consortium of 16 national archives, libraries, research institutions and leading IT companies.

Further Information

For more information about Planets visit: **http://www.planets-project.eu**

You can email your questions to us at:
info@planets-project.eu

The Open Planets Foundation (OPF) builds on the investment made in the Planets project. It will sustain the results of this investment and further develop and coordinate development of the capabilities that its members require. It will provide services, knowledge, methods and tools to its members and the broader community.

For further information visit: **http://www.openplanetsfoundation.org**