A Case Study on Performing a Complex File-format Migration Experiment using the Planets Testbed

Sven Schlarb, Edith Michaeler, Max Kaiser; Austrian National Library; Vienna, Austria; Andrew Lindley; Austrian Institute of Technology; Vienna, Austria; Brian Aitken, Seamus Ross; HATII at the University of Glasgow; Glasgow, Scotland; Andrew Jackson; The British Library; Boston Spa, England

Abstract

Digital Preservation deals with the long-term storage, access and maintenance of digital objects. Preservation experts are faced with an almost overwhelming variety of preservation actions and tools but, as individuals, lack the time and resources required to build the evidence base that an informed decision process requires. As one of the core developments of the Planets project, the Planets Testbed offers a scientific research environment to perform and document experiments for digital long term preservation. This paper provides an overview on the Planets Testbed and provides a case study demonstrating how the Testbed can be used to perform complex experiments using Planets preservation services, large datasets and the automated evaluation of results.

1. Introduction

Memory institutions face the challenge of preserving large collections of electronic resources both online and offline. The quantity of digital objects and the range of file formats encountered are increasing rapidly and the long-term preservation of such complex collections requires a detailed understanding of digital media, including information pertaining to the characteristics of file formats, knowledge about valid metadata extraction methods, strategies relating to effective and non-invasive format migration, and strategies to present digital items in an emulated environment to ensure an authentic user experience. To address these issues, time and labour-intensive software tests and feasibility studies may be performed by different institutions in parallel, but without the publication of results the valuable outcomes cannot broaden the knowledge of the community.

The Planets Project (Preservation and Long-term Access through NETworked Services) has created a software suite capable of addressing the core digital preservation challenges that libraries, archives and the digital preservation community are currently facing. Based on a common platform, the Planets Interoperability Framework, it offers shared access to services and tools for planning and performing experiments in the domain of long-term preservation. As a major software outcome of the Planets project, the Planets Testbed is a web-based application that offers registered users the possibility of executing, documenting and publishing preservation experiments, thus enabling the dissemination and sharing of results with the wider community.

The Planets Testbed offers a dedicated online instance which gives access to Planets and third party developed preservation services in a controlled research environment, thus enabling the evaluation of services against defined evaluation criteria.

In this paper we show how the Planets Testbed can be used create and evaluate complex software based workflows using different migration services and a sample dataset comprising 750 TIFF image files from the collections of the Austrian National Library.

2. An Overview of the Planets Testbed

The Planets Testbed[1] provides a simple way to assess the behaviour of tools and objects within a controlled and objective environment. A highly configurable six-step process standardises the way in which different experiments can be performed and ensures that each experiment is executed under comparable conditions. This facilitates analysis and helps guarantee that all results gathered can be reproduced over time. The Testbed also provides a collection of annotated files, the Testbed Corpora, offering realistic test datasets for evaluating preservation services. Containing several thousand files of different origins and formats, the corpora are annotated with both descriptive metadata and additional "property" metadata that allows the comparison of objects through the XCDL comparator[2]. The Austrian National Library Corpus featured in this case study is part of the Testbed Corpora and contains a large set of genuine digital media content generated during a digitisation project.

2.1 Experiment Types

The Testbed features three major experiment types, each of which can be customised to the needs of individual experimenters and their organisations:

- Migration: With this experiment type one or more digital objects can be passed through a migration service which will attempt to transform each digital object into a new format that the experimenter may specify. In addition to this, experimenters may optionally run characterisation services on the input and output files after execution to compare the results.
- View in Emulator: This experiment type allows the experimenter to pass one or more digital objects to an emulator, which will then open within the experimenter's web browser, enabling the digital objects to be manipulated within the emulated environment.
- Execute Preservation Plan: With this experiment type the experimenter may take an existing executable preservation plan that has been generated by a preservation planning tool such as the Planets Preservation Planning Tool Plato [14] and test the plan within the Testbed environment.

The presented case study features an executable preservation plan experiment type, containing a chained sequence of migration and characterisation services that has been designed as a load test migration experiment. The experiment was set up to transform large numbers of digital objects while providing a focus on the selected tools' speed and stability.

2.2 Reproducibility of Results

A core Testbed principle is the community-centred approach of building a continuously growing shared-knowledge base about preservation services and their underlying tools. For this reason all Testbed results are available to the community and can be shared with other users for online and further offline analysis. In addition, experiment setups can be easily copied, modified and rerun using the Testbed's 'save as' functionality. The experiments that form the basis of the case study presented within this article, including setup, data and a complete set of offline evaluation information can be accessed at through the Testbed at:

http://testbed.planets-project.eu/testbed/exp/exp_stage5.faces?eid=343

3. The Case Study

To illustrate the Testbed's aptitude for digital long term preservation experimentation, the project group created a case study to investigate Testbed performance and an individual collection holder's needs. In this case a content-holding institution is evaluating the option to migrate an image collection available in uncompressed TIFF to JPEG 2000 (JP2) in order to save storage space and to take advantage of the advanced features of the JP2 file format in their long-term preservation environment. For the sample scenario the JasPer [9] and Kakadu [10] image codecs were selected for comparison.

3.1 Experiment Objectives

The case study was set up to illustrate three different aims: Firstly to show how the Planets Testbed may be used to perform a complex file format conversion experiment; secondly to demonstrate the impact of the performance of services and tools on images in a complex migration workflow; and thirdly to assess the applicability and reliability of software tools within the Planets execution environment.

Related to this case study, the typical Testbed approach to such a scenario is presented, focusing on concrete preservation experiments. For example, services based on JP2 conversion tools are used, and studies related to the JP2 file format and tools are presented in [3], [4], and [5] while another example of showing the use of Planets tools for long-term preservation with a very similar focus is presented by [6]. The experiment documented in this paper presents results about the following preservation related aspects:

- The reliability of services, specifically indicating if the services involved in an automated software-based workflow can be successfully applied to a large sample dataset.
- Service and tool execution performance, defined through the measurement of execution time and throughput (megabytes per second) for each of the services involved in an automated software-based workflow.
- Comparison of the JP2 encoding and decoding performance of services based on JasPer and Kakadu.
- Automated Quality Assessment in terms of deviation measurements between the original and the resulting TIFF files provided by the XCDL Compare service.

The experiment workflow was designed to show a typical application scenario for assessing the applicability of selected services for a migration strategy that has previously been analysed within Planets using the preservation planning software Plato[14]. Although the basic assumptions of the experiment are based on concrete possible needs, the outcomes are not intended to provide a basis for strategic decisions, nor do they intend to claim universally valid conclusions related to the suitability of tools or file formats from the long-term preservation perspective.

3.2 Meeting Experiment Objectives in the Testbed

In order to reliably determine service and tool execution performance, the Planets Testbed hosts preservation services in a controlled environment. In order to fully understand the factors that dictate the speed of execution, it is necessary to record information about the execution process and the host environment. For example, the majority of hosted web services are written in Java and execute inside a Java Virtual Machine (JVM). This implies that the service execution process will be competing with the other threads in the JVM, such as the garbage collection thread, and therefore complete isolation of the services cannot be achieved. Therefore the total time for a service to complete may be a consequence of resource contention on the server, rather than the actual service processing. Fortunately, the Java language also provides a number of mechanisms that allow more detailed measurements to be performed during execution, including:

- Service execution time: Total run time of the service, i.e. the time when service processing terminated minus the time when service processing was initiated.
- Wall-clock time: The elapsed time taken to complete the task, as would be measured by a clock on the wall.
- Wall-clock load time: The elapsed time taken to transfer the content to the service and load it into memory, as would be measured by a clock on the wall.
- CPU time: Processing time devoted to the task.
- User time: The elapsed time taken to execute user code, as opposed to system or input/output operations.
- Tool execution time: The tool execution time that is measured for tools (e.g. command line tools). The tool execution time is measured as the run time of the java thread which executes the tool.

These measurements allow a precise analysis of the service performance, and combined with information about the host platform, allow the computational and I/O requirements of a particular preservation plan to be explored. The feasibility of the plan can then be estimated by predicting the computational resources required to implement a particular plan on a particular collection.

3.3 Services Used in the Experiment

The case study presented in this paper has involved different file format migration services, such as the JP2 codec tools JasPer and Kakadu, and image conversion tools such as Gimp[11] and Sanselan[12]. Characterisation tools, such as JHove[13] and the XC*L software suite[7] are used for automated quality assurance and experiment evaluation.

Web services offering access to tools through the Planets service registry usually provide access to a small subset of the sometimes very rich functionality of the tools deemed to be relevant for performing long-term preservation operations. The advantage of this approach is that it offers a "long term preservation perspective" on the broad variety of software available for this purpose. In this sense, the service parameters

accessible and configurable through the Testbed web interface represent a pre-selection of parameters relevant for performing long-term preservation experiments.

As an example of the service configuration which is used in the experiment scenario, Table 1 shows the KakaduCompress service parameters for encoding TIFF images to JP2:

Parameter Name	Value	Description
rate	1.0	Comma separated positive floating point numbers indicating the compression rate. For example: 1.0 means irreversible compression to 1 bit/sample
reversible	yes	Indicates whether the compression is reversible (lossless) or not.
layers	5	Number of wavelet decomposition levels, or stages.
levels	5	Embedded quality layers
tiles	1024,1	After color transformation, the image is split into so-called tiles, rectangular regions of the image that are transformed and encoded separately.
cblk	64,64	Codeblocks are used to partition the image for processing and make it possible to access portions of the datastream corresponding to sub-regions of the image.
order	LRCP	Indicates the progression order. The four character identifiers have the following interpretation: L=layer; R=resolution; C=component; P=position.

Table 1: KakaduCompress Service Parameters for PNM to JP2 Encoding

3.4 The eXtensible Characterisation Language Suite

Automated quality assurance of the migration outcomes of the experiment workflow are managed through the eXtensible Characterisation Language Suite which has been extended within the Planets project.[2] XCDL provides an abstract representation of a digital object's content, and XCEL describes the way in which the properties of a particular digital object format can be extracted and mapped into a format neutral XCDL model.

A key objective of the XCDL Comparator is the property-specific definition of metrics and their underlying algorithms to identify degrees of equality. Within this experiment, similarity is evaluated in terms of the root mean squared error (RMSE) [7], a widely used measure for deviation, as shown in Figure 1 below. It is obtained by calculating the square root of the average of the squared differences of single values on two sets of normalised data taken from the input and output XCDL files created as part of the experiment's workflow. Beyond the presented RMSE metric the XCDL Comparator service was also configured to automatically compare the properties image height, width, bits-per-sample and alpha-transparency.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_{Ai} - x_{Bi})^2}{n}}$$

Figure 1:Root Mean Squared Error

4. The Experiment Scenario

The experiment scenario comprises of the selection of input data, in this case a sample dataset of TIFF files, the creation of workflows by combining services such as JasPer, the experiment setup consisting of a combination of these workflows together with a comparison methodology, and the evaluation of experiment results.

In this paper, the term workflow refers to a software-based execution of a sequence of tasks, without human interaction, which is executable on the Planets integrated environment [8]. In the narrower sense it is a sequence of tasks operating on image data, either performing an image file format conversion, extracting metadata information, or comparing image characteristics. The output of one service forms the input of the next service in the processing chain, and for each of the JasPer and Kakadu services a round trip migration pathway is implemented as a workflow. "Round trip" in this case means that each workflow starts with the original TIFF, performs a JP2 encoding, and finally converts the image back to TIFF with PNM and PPM as intermediate temporary file formats, as Figure 2 demonstrates.

The experiment setup is the combination of the two workflows described above, together with a comparison methodology (e.g. comparing JP2 encoding performance by comparing the throughput in terms of Megabytes/Milliseconds). Figure 2 represents the experiment set up, illustrating the use of two separate migration pathways. The first path uses JasPer and the second uses Kakadu for encoding and decoding JP2. As previously mentioned, both workflows feature a "round trip" conversion process, converting the original TIFF to JP2 and then back once more to TIFF. As part of each single workflow, the original TIFF file and the resulting TIFF file are compared to each other using the XCLCompare service in order to determine success and reliability of the migration pathway. Additionally, encoding and decoding performance comparison between JasPer and Kakadu (represented by dotted lines in Figure 2) is undertaken by the evaluation of execution logs.

The plain ASCII text image formats PNM and PPM have been chosen as the common base format enabling encoding and decoding performance comparison of JasPer and Kakadu. It must be noted that these formats are not recommended to be used in a production environment but simply have been chosen in order to reach the "round trip" experiment setup and in order to keep the experiment setup simple for demonstration purposes.

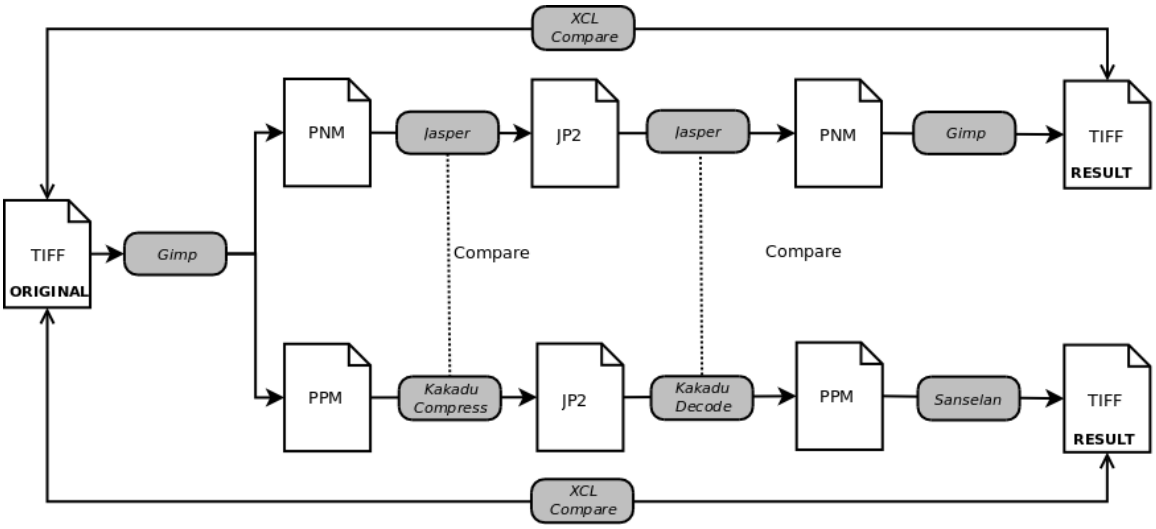

Figure 2: Experiment Set-up consisting of two workflows converting TIFF to JP2 and vice versa using JasPer and Kakadu as JP2 codecs

5. Evaluation of Results

In this section the results of the sample experiment are presented. Evaluation includes an outline of the overall performance measurements as mean values relating to the 750 item dataset and a presentation of the results gathered through the automated image comparison methods for determining the success and reliability of the migration strategy.

In order for the results to be meaningful it is necessary to provide some information about the hardware and software environment through which the Testbed experiments were performed. The Testbed server, located at HATII at the University of Glasgow, comprises of a 2.5Ghz Intel Quad Core Xeon E5420 with 16GB RAM. The server runs 64-bit Ubuntu 8.04 through which an 8GB Java Virtual Machine is configured.

5.1 Tool and Service Performance

It is not possible to include detailed execution results in this paper, however these are available through the Testbed via the URL presented above. Following the execution of a range of experiments on the dataset the following general observations were noted:

- There is an overhead when using the JHove identification service when processing smaller files. This can be observed by looking at the mean service execution time when comparing the two paths, starting from 69.84 ms/MB on an average file size of 24 MB in the JasPer path and reaching up to 986.75 ms/MB on an average file size of 0.46 MB in the Kakadu experiment, which is therefore slower by a factor of 14.
- The XCDL Extractor was observed to be robust in this respect and was able to gain a 12.8% speed advantage.
- When processing the compression of JP2, Kakadu gained a clear speed advantage over JasPer by a factor of 2.53 with respect to tool execution time.
- The decoding from JP2 to PNM/PPM takes roughly twice as long than encoding from PNM/PPM to JP2 for the JasPer service and 3.5 times longer for the Kakadu service.
- When encoding digital objects with the selected parameters the JP2 encoded objects created by the Kakadu service are a 35% smaller compared to those created by the JasPer service, although the intermediate input images are approximately the same size for both pathways.
- The best service implementation measured in terms of use of computing resources by the service (including web-service data transmission, code compilation, etc.) in comparison to the actual tool execution time is JasPer in the migration pathway PNM to JP2. In this instance there is an additional service overhead of 30% compared to a 100% overhead using the Kakadu service. Unfortunately Sanselan and the XCDL migration tools do not record sufficiently detailed information about the underlying processes to make a statement.

It should be noted that the comparison of Sanselan and Gimp in the final step of the roundtrip migration is not adequate as the Kakadu path was configured to produce smaller output files on which the lossless Lempel-Ziv-Welch (LZW) compression algorithm was applied. This approach was chosen in order to compile meaningful measurements for the comparison of the two experiment paths in terms of RMSE and automated QA.

5.2 Automatic Quality Assurance

In order to automatically determine the success and reliability of image migration, specific digital object properties, for example normalised data, image height, image width, bits per sample and the alpha transparency of the original TIFF file and the migrated result file have been compared by means of the Planets XCDL Compare service. On the basis of measurements, such as the RMSE measurement mentioned above, this service provides some indicators about the success and reliability of the execution. The overall result does not allow conclusions to be drawn regarding specific reasons for failures, lack of performance, or other

deficiencies of the workflow, but instead provides an overall indicator about the migration path.

These results indicate that the generated output files from the roundtrip migrations were completely equal in terms of image height, image width, and bits-per-sample compared to the original input files for both the JasPer and Kakadu pathway of the experiment.

Figure 3 below shows the 683 successful comparison results ordered by the RMSE value in ascending order for the JasPer path. In nine cases no distance measure could be computed. Most of the RMSE values are zero or very close to zero (first quartile=0; median=0.066; third quartile=4.07) which is an indicator of successful migration. On the other hand, a relatively large number of results do have high RMSE values (arithmetic mean=12.69). This indicates major differences from the corresponding norm data, and to understand the exact reasons for this further analysis of these objects can be carried out through the detailed execution log or by downloading the corresponding items and performing a separate Testbed experiment. For example, for the highest measured RMSE value of 94.72 within this experiment path, the files appear to be visually equal, but the reason for the high RMSE value is that the original TIFF image had an embedded color profile while the color space is converted to RGB in the resulting TIFF file.

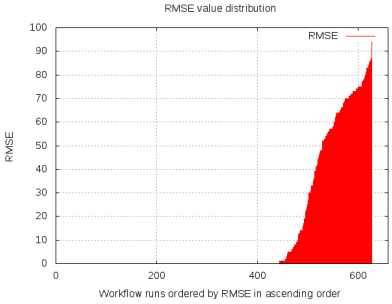

Figure 3: Workflow runs of the JasPer experiment path ordered by RMSE in ascending order

Similarly, it is possible to analyse the results for the Kakadu experiment path, which demonstrates that there were 477 successful RMSE measurements, as shown in Figure 4 below. It is striking that compared to the JasPer pathway, the 621 RMSE values are very high (arithmetic mean=51.19) even in the first quartile with 39.46 and a median of 45.56. This might be an indicator that the Kakadu execution is more robust because it has a lower number of problematic cases with extremely high RMSE values in the sample data set. However, as previously stated, in order to draw reliable conclusions, further manual analysis of the items where the measurement failed and of the items with very high RMSE values would be required.

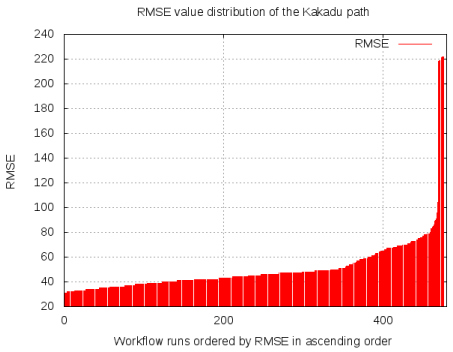

Figure 4: Workflow runs of the Kakadu experiment path ordered by RMSE in ascending order

6. Conclusion

The selected services were found to be robust and capable of performing the required migration steps with a success ratio of 98.5 percent within the Planets Testbed software and hardware environment, and therefore the Testbed proved to be a solid environment for executing and evaluating the experiment described in this paper.

The large overall data volume indicates that the Planets Workflow Execution Engine [8] and the Testbed environment are able to process large data sets. The sample dataset consisting of 750 TIFF image scans with a total volume of 17.4GB and an average file size of 23.7 MB used in this experiment was successfully processed over several experiments within the Testbed. The overall volume of data created by both the Kakadu and the JasPer pathways was 252 GB and as an overall result, we can indicate that the average workflow execution time to process one megabyte of data over all services was 42.1 seconds for the Kakadu path and 13.3 seconds for the JasPer path.

Using the XCDL Comparator service for automated quality assurance provided some indicators about the overall success of the migration pathways used in the experiment scenario presented in this paper. However, in order to provide better reporting of the quality of migration, more indicators have to be taken into consideration, and detailed reporting regarding problematic cases needs still significant improvement.

Ackowledgements

The Planets Testbed Research and Development work presented here is partially supported by European Community under the Information Society Technologies (IST) Programme of the 6th FP for RTD - Project IST-033789. As the Planets project is drawing to a close, steps have been taken to ensure that the Testbed and the other Planets software and documentation will be maintained. The project consortium has worked to establish the Open Planets Foundation, a membership-based not-for-profit company dedicated to the long-term sustainability of the work that the Planets project began.

References

[1] Andrew Lindley, Andrew Jackson, Brian Aitken: A Collaborative Research Environment for Digital Preservation - the Planets Testbed, 1st International Workshop on Collaboration tools for Preservation of Environment and Cultural Heritage at IEEE WETICE, 2010, submitted for publication, URL http://planets-project.ait.ac.at/publications/PlanetsTestbed_COPECH_08032010.pdf

[2] Manfred Thaller (Ed.): The eXtensible Characterisation Languages – XCL, Verlag Dr. Kova , ISBN 978-3-8300-4766-7, 2009.

[3] Farzad Ebrahimi, Matthieu Chamik, Stefan Winkler: JPEG vs. JPEG2000: An Objective Comparison of Image Encoding Quality, 2004, URL http://stefan.winklerbros.net/Publications/adip2004.pdf.

[4] Paolo Buonora und Franco Liberati: A Format for Digital Preservation of Images, 2008, URL http://www.dlib.org/dlib/july08/buonora/07buonora.html.

[5] Robèrt Gillesse, Judith Rog, Astrid Verheusen: Alternative File Formats for Storing Master Images of Digitisation Projects, 2008, URLhttp://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/alternative_file_formats_for_storing_masters_2_1.pdf

[6] Hannes Kulovits, Andreas Rauber, Anna Kugler, Markus Brantl, Tobias Beinert, Astrid Schoger: From TIFF to JPEG 2000? Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, D-Lib Magazine November 2009, Volume 15, Number 11/12, 2008, URL http://publik.tuwien.ac.at/files/PubDat_182423.pdf.

[7] In the context of the experiment presented here, the XCL software suite refers mainly to the XCDL Extractor and XCDL Comparator tools. Regarding the XC*L languages see Manfred Thaller et al.: Planets deliverable WP PC/2-D12, PC/2-D13, PC/4-D7, eXtensible Characterisation Language Suite, 2009

[8] Rainer Schmidt, Ross King, Andrew Jackson, Carl Wilson, Fabian Steeg, Peter Melms: A Programming Model and Framework for Distributed Preservation Workflows, in Proc. International Conference on Preservation of Digital Objects (iPRES), 2009

[9] JasPer Transcoder Version 1.900.1 specialised for JPG to JP2 (JPEG2000) encoding and vice versa was used. Copyright (c) 1999-2000 Image Power, Inc. and the University of British Columbia. See http://www.ece.uvic.ca/~mdadams/jasper

[10] Kakadu version 6.2.1 command line application which shows the potential of the JPEG 2000 Developers' Toolkit is used. See http://www.kakadusoftware.com

[11] Gimp 2.6 - GNU Image Manipulation Program, see http://www.gimp.org

[12] Sanselan: a Pure-Java Image Library, see http://commons.apache.org/sanselan

[13] JHOVE - JSTOR/Harvard Object Validation Environment, see http://hul.harvard.edu/jhove

[14] Planets Preservation Planning Tool, see http://www.ifs.tuwien.ac.at/dp/plato

Author Biographies

Sven Schlarb holds a PhD in Humanities Computer Science from the University of Cologne. Before joining the Austrian National Library, where he is participating in the EU-funded projects Planets and IMPACT, he worked as a software engineer in Cologne and Madrid and as support consultant at SAP in Madrid

Edith Michaeler studied History and Political Sciences in Vienna and Paris and holds a Masters degree in History (2004). She is also trained in journalism and PR (Diploma 2006). At ONB Edith is working in the EU-funded project Planets as deputy sub-project lead of the Testbed sub-project, as well as for in dissemination, training and sustainability planning in digital preservation.

Max Kaiser joined the Austrian National Library in 2000 and is currently Head of the Research and Development Department. He has many years of experience in the field of digital libraries, digital preservation and digitisation. In Planets he has been member of the Scientific Board and lead of the Testbed sub-project. He is coordinator of the EU-funded EuropeanaConnect project which is implementing essential components for Europeana.

Andrew Lindley joined the Austrian Institute of Technology in 2006 and is part of the Safety & Security Department's digital preservation team. He has a master degree in economics and computer science and has recently worked in the EU-funded project Planets where he has was closely involved in development and design of the Planets software.

Brian Aitken joined the HATII in 2001 as a Systems Developer. He has worked on a number of successful projects, most recently EU-funded Planets, for which he is leading the development of the Testbed, and EU-funded SHAMAN. Previously he has managed and developed online tools and content management systems for the Digital Curation Centre, DigitalPreservationEurope and DigiCULT, and for several successful digitisation projects.

Seamus Ross is Dean and Professor in the Faculty of Information University of Toronto. Formerly, he was Professor of Humanities Informatics and Digital Curation and Founding Director of HATII (Humanities Advanced Technology and Information Institute) (1997-2009) at the University of Glasgow.

Andrew Jackson has a background in computational physics and software architecture, including many years of experience designing algorithms for very large scale computation and data management. As a Digital Preservation Architect at the British Library, he has been closely involved with the development of the technical architecture of the Planets software, the establishment of the Open Planets Foundation, and with integrating those products with the British Library preservation systems