ARTICLES

D-Lib Magazine
November/December 2009

Volume 15 Number 11/12

ISSN 1082-9873

From TIFF to JPEG 2000?

Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings

Hannes Kulovits, Andreas Rauber
Vienna University of Technology
Vienna, Austria
{kulovits, rauber}@ifs.tuwien.ac.at

Anna Kugler, Markus Brantl, Tobias Beinert, Astrid Schoger
Bavarian State Library
Munich, Germany
{brantl, beinert, kugler, schoger}@bsb-muenchen.de

Introduction

Studies and user reports claim JPEG 2000 to be – or at least will become – the next archiving format for digital images [1]. The format offers new possibilities, such as streaming, and reduces storage consumption through lossless and lossy compression [2]. Another often claimed advantage of JPEG 2000 is that the master image can possibly serve as the access copy as well, and thus replace derived compressed, low resolution access copies. The National Library of the Netherlands (KB-NL) evaluated the suitability of alternative file formats such as JPEG 2000 to their currently used format uncompressed TIFF. The four aspects, required storage capacity, image quality, long-term sustainability and functionality were analysed and JPEG 2000 is recommended as future archive format [3]. The British Library recently moved forward to migrate their 80-terabyte newspaper collection from TIFF to JPEG 2000 [4] and the Wellcome Library announced they will use JPEG 2000 for their upcoming digitization projects [16].

Having the advantages of JPEG 2000 in mind, the Bavarian State Library (Bayerische Staatsbibliothek, BSB [5]) also considered the option of migrating from TIFF to JPEG 2000 as the archive format for digitized images of rare books. BSB aims at digitizing its complete collection of manuscripts and rare books, applying high standards and policies (e.g., referring to compression, resolution, and colour management) that result in considerable image sizes of the TIFF-master copies (which, for example, in the 15th century incunabula project average 100 MB per page).

In order to find out whether TIFF or JPEG 2000 would be a more suitable archival master format, the BSB, together with the Vienna University of Technology, created a preservation plan for a representative collection of digitized 16th century printings. The goal of the project was to evaluate possible strategies for migration from TIFF to JPEG 2000 using lossless compression, including the alternative of keeping the status quo. The current preservation plan documents the resulting decision, taking into consideration the institution's preservation policies, legal obligations, organizational and technical constraints, requirements, and preservation goals, as well as the capabilities of the tested tools.

Preservation Planning is a process that depends very much on an institution's individual policies and requirements in its day-to-day work. Consequently, the design of one plan can differ considerably from the plan of another institution, even one with a similar collection.

The creation of the preservation plan described in this article follows the Planets Preservation Planning Workflow, which is described in detail in the *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)* [6]. The plan was created using the open source planning tool Plato, which is available to the public at <http://www.ifs.tuwien.ac.at/dp/plato>.

The Bavarian State Library: Digitization and long-term preservation

The Bavarian State Library (BSB) is a universal and research library of worldwide importance not only due to the size of its collection but also to the uniqueness of its inventory of handwritings, incunabula and historical printings. The BSB holds almost 10 million volumes and more than 50,000 current periodicals in both analogue and digital form. As the central state and archive library for Bavaria, as well as part of Germany's virtual national library (together with the Berlin State Library and the German National Library), the BSB is responsible moreover for collecting, indexing, archiving and making available analogue and digital information resources.

The BSB strategy is to digitize its entire out-of-copyright holdings within the next five years and to make them accessible free of charge via the Internet. BSB pursues this strategy through mass digitization projects, e.g., those funded by the German Research Foundation [7] (Deutsche Forschungsgemeinschaft, DFG) and BSB's public-private partnership with Google books: BSB receives a copy of each digital object produced by Google. Consequently, the BSB is among the cultural heritage institutions in Germany with the fastest growth of digital objects in its archive (by August 2009, 153 TB). As an active player in the field of long-term preservation, BSB cares not only about continuous access to its preserved information in the present but also into the future.

The responsible department within BSB is the Munich Digitization Centre [8] (Münchener Digitalisierungszentrum, MDZ), which has been engaged as an innovation centre for digital information services in the world of digital libraries since 1997. The MDZ's main tasks at the moment are the mass-digitization of printed material, development and maintenance of subject portals as well as user-oriented services, and the long-term preservation of all kinds of digital objects within its collection profile. All digital objects, including born-digital materials, are stored at the Leibniz Supercomputing Centre [9] (Leibniz Rechenzentrum, LRZ). By August 2009, the total number of digital files in archival storage there had reached 212.252.017.

The digitization policy of the MDZ is to produce and store high-quality digital images. All analogue material from the 6th century up to the 18th century (printings, manuscripts, rare books, historical maps, etc.) underlies defined digitization standards:

- TIFF uncompressed
- High resolution (300-600 ppi, in relation to the original size of the document)
- 24-bit colour depth
- ICC profiles embedded for colour fidelity
- No postprocessing of the images

The MDZ preserves the original image files to guarantee their integrity, authenticity and availability over the long run, to ensure protection of investment and to allow the re-use of the digital objects in the sense of cross-media publishing (e.g., Internet publishing, document delivery, ebooks-on-demand, printing in catalogues or facsimile editions).

National and international standards are applied and best practices acknowledged for digitization as well as for long-term preservation. The continuous evaluation and improvement of the quality of its technological and organizational infrastructures and processes are also part of MDZ's policy. In a current project[1] MDZ checks and improves the trustworthiness of its digital archive according to the nestor criteria catalogue [10]. Designing and testing new digital

preservation strategies, and thus gaining the necessary competence to apply innovative methods and tools for preservation planning, is one part of the project.

Preservation planning and Plato

The Planets project has developed a systematic approach for evaluating potential alternatives for preservation actions and building thoroughly defined, accountable preservation plans for keeping digital content alive over time. In this approach, preservation planners empirically evaluate potential action components in a controlled setting and select the most suitable one with respect to the particular requirements of their institution [11]. The procedure is independent of the solutions considered; it can be applied for any class of strategy, be it migration or emulation or different approaches. The method follows a variation of utility analysis to support multi-criteria decision making procedures in digital preservation planning. The selection procedure leads to well-documented, well-argued and transparent decisions that can be reproduced and revisited at a later point in time. The planning tool Plato [12] supports, documents, and automates the decision procedure and produces a preservation plan as result. The Planets project has defined a preservation plan as follows [11].

> "A preservation plan defines a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records (called collection). The Preservation Plan takes into account the preservation policies, legal obligations, organizational and technical constraints, user requirements and preservation goals and describes the preservation context, the evaluated preservation strategies and the resulting decision for one strategy, including the reasoning for the decision. It also specifies a series of steps or actions (called preservation action plan) along with responsibilities and rules and conditions for execution on the collection. Provided that the actions and their deployment as well as the technical environment allow it, this action plan is an executable workflow definition."

The planning workflow that leads to the preservation plan prescribes four phases:

1. Define requirements
2. Evaluate alternatives
3. Analyze results
4. Build preservation plan

The first phase '*Define requirements*' documents constraints and influence factors on potential preservation strategies. It then continues with a thorough description of the collection and the chosen sample objects from that collection, and concludes with the definition of the complete set of requirements. At the end of this phase the planner has a detailed and exact understanding of the collection and the preservation goals. The elicitation of the institution's requirements is the core activity in the planning workflow and is vital, as the requirements co-determine the optimal preservation action within the institution's context. Not all university libraries or national libraries, for instance, share the same objectives.

The second phase '*Evaluate alternatives*' starts with discovering potential preservation actions (alternatives) that are then evaluated in a quantitative way. Controlled experiments are carried out, applying the alternatives to the defined sample objects and analysing the outcomes with respect to the requirements. The result of this phase is an evidence base that underlies all decisions to be taken in the subsequent phases.

In the third phase '*Analyze results*', the results of the experiments are analysed and aggregated. The result of this phase is a ranked list of alternatives wherein the alternative with the highest performance value presents the recommended preservation action.

In the final phase '*Build preservation plan*', based on the recommended preservation action, a preservation plan is created that corresponds to the Develop Packaging Designs & Migration Plans functionality in the OAIS model [13].

The planning process

With the help of the preservation-planning tool Plato, we followed the previously described workflow. In this article, we describe the work we've done and the sample objects we've chosen, as well as discuss the requirements, considered alternatives, experiment settings, and results of our work.

Description of the collection

We created the preservation plan for one of the largest digital collections within the library. The collection was digitized in the course of a project funded by the DFG between 2007 and 2009. The project aimed at digitizing, indexing, delivering and preserving all 16th century prints (1518-1600) in the German language held by BSB. It was initiated within the context of the DFG campaign as the first mass digitization project of German cultural heritage.

About 60 % of the prints were scanned automatically in-house with Scan-Robots® in order to achieve an efficient throughput under BSB's strict conservational requirements and with a high-quality digitization output of early prints. During the project about 21.000 volumes with more than four million pages (around 72 TB) were digitised. Nevertheless, this collection displays a very homogenous compilation of digitized images. All master files are TIFF uncompressed (version 6.0) with a resolution of 300 ppi. The prints were scanned in colour with 24-bit colour depth. To gain the best colour fidelity using any output device (e.g., for cross-media publishing), along with every image we store an adequate ICC-colour profile. Copies for web-access are JPEG-files. In addition, sharpness and colour target are scanned and displayed on one selected page of each digitized volume, in order to enable additional visual examination. All digitized images of the project are run through different quality control procedures after the scanning process, and only if the images comply with MDZ standards, are the files transferred into the archival storage at LRZ.

During the semi-automated production process, the bibliographic, technical, administrative and structural information of each volume is stored with our electronic publishing framework ZEND (Zentrale Erfassungs- und Nachweisdatenbank), which is based on Open-Source Software, in a separate XML file based on TEI P5, which we use for the correct access of the images (physical and logical structure of the images are mostly not congruent). The technical metadata of the images are always stored representatively for the whole volume on the 11th page of each volume. When substantial progress has been made in the development of gothic font recognition, we plan to integrate automatic OCR-extraction (reusing the archived digital master files). As a partner in the IMPACT project, we expect a step forward will be made in the future in OCR development for 16th century prints [14].

Based on an analysis of the collection's profile, we chose six sample objects stratified across file size and content. The smallest image has a size of around 11 MB and the largest one has an image size around 30 MB. All of them have TIFF properties embedded, such as the artist, a timestamp and the software used for scanning.

The requirements concerning the collection

At the heart of the preservation-planning approach are the requirements the optimal preservation solution must meet. We conducted a requirements elicitation workshop that involved stakeholders from the head of the digital library and digitization services, digitization and preservation experts, and employees from the library and the LRZ. We followed a top-down approach and started with the high-level objectives: *Technical characteristics* (*of the target format*), *Object characteristics*, *Costs*, and *Process characteristics*. Figure 1 illustrates the requirements tree at which we finally arrived, based on those high-level objectives and broken down to measurable criteria. The entire process of defining the requirements is influenced by numerous factors, including standards, legal constraints, and user and organizational requirements and policies. Some of the requirements are:

> **Maintain high quality**, e.g., Compression. Does the file format to which we want to migrate support lossless compression?
>
> **Maintain authenticity**:

- *Image identical*. Did any of the pixels change during migration? We evaluated this using ImageMagick's and GraphicsMagick's "compare" command to check the original image against the migrated one. Both tools enable measurement of the absolute number of different pixels. Since both tools are also amongst the main protagonists for the migration that we actually want to evaluate, we decided that only if both tools purport the images to be identical would we assume that the migration tool had fulfilled this requirement, i.e., the images are identical. This requirement is a drop out criterion in the preservation plan, meaning that a preservation solution must fulfill it to be further considered in the evaluation.
- *Resolution identical*. Did the migration tool preserve the resolution of the image?
- *Colour profile identical*. Was the tool able to preserve the embedded colour profile?
- *Metadata*. Did the tool maintain embedded metadata (e.g., provenance scanner, date, ...)

Allow creation of full-text: *OCR possible*. Does the OCR software deployed at BSB support the new file format?

Have the quality of ubiquity: How well adopted is the new file format? Do web browsers support it, and are there enough tools available?

Reduce storage costs: We experience a continuous decrease of storage costs at LRZ, and consequently, storage costs play a minor role for us.

We created the requirements tree during the elicitation workshop using the open source tool Freemind. The resulting tree with the requirements and measurement scales can be uploaded to Plato and used for subsequent steps.

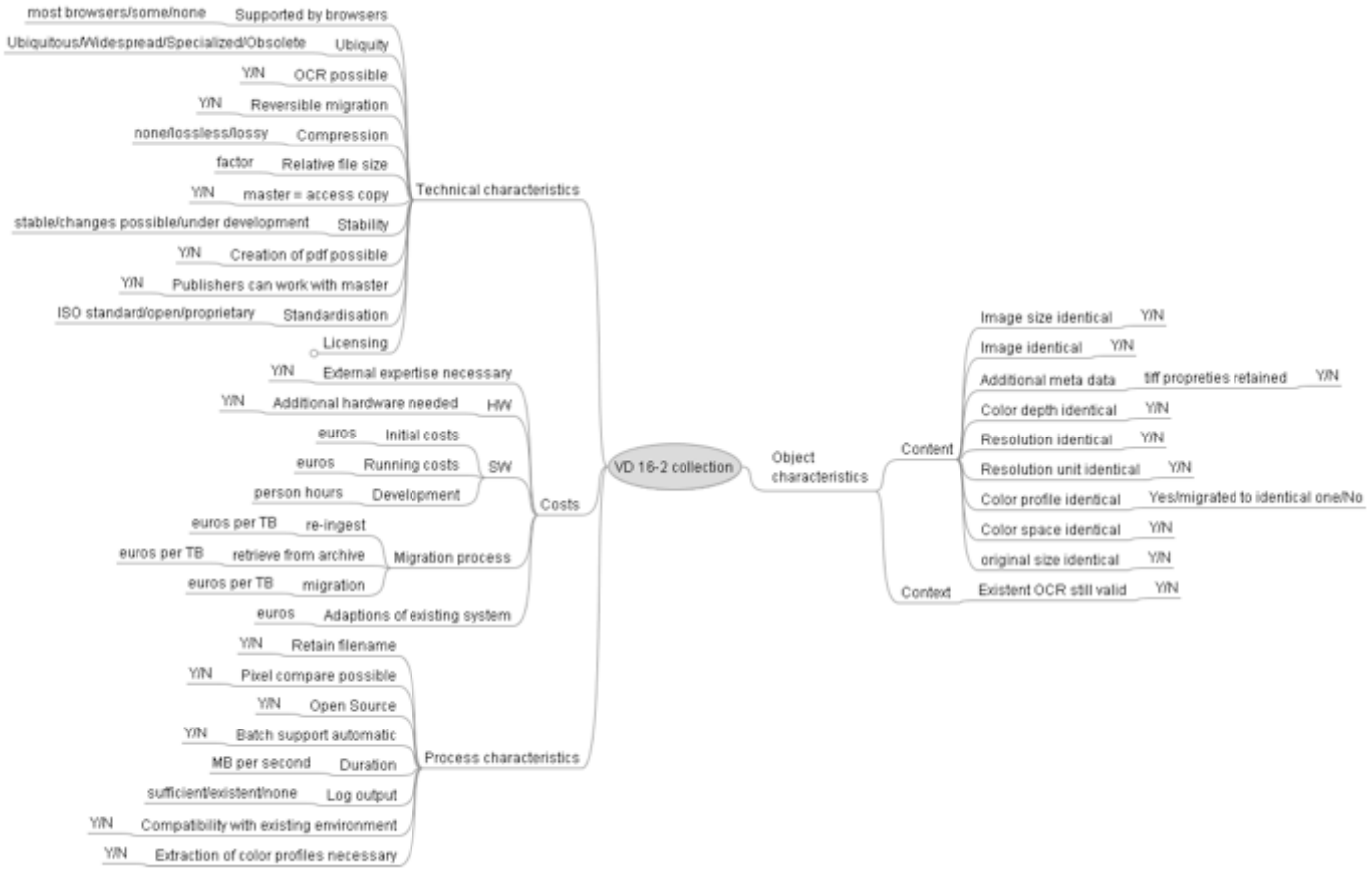

Figure 1: Requirements tree
(For a larger view of this image, click here.)

Potential preservation actions

Following the BSB policies we considered four open source tools that are able to perform TIFF to JPEG 2000 migration using lossless compression. We also included 'Keep status quo' in the evaluation, since at present there is no particular risk associated with TIFF uncompressed. Therefore leaving the collection untouched is a viable option. Table 1 gives a list of the five alternatives with the parameters we used for the actual migration.

Table 1: Alternatives considered for evaluation using the four listed software tools

Alternative	Description	Parameters
Keep status quo	Leave the images in TIFF v6	
ImageMagick TIFF to JP2000	Version: 6.5.3-5 2009-06-11 Q16	`-compress Lossless -quality 100`
GraphicsMagick TIFF to JP2000	Version: 1.3.6 2009-07-25 Q8	`-compress Lossless -quality 100`
Kakadu TIFF to JP2000	Version 6.1	`Creversible=yes -rate -,1,0.5,0.25 Clevels=5`
GeoJasper TIFF to JP2000	Version: 1.3.1	`-T jp2`

Experiment setting and evaluation of outcome

The BSB experiments have been conducted on a Windows XP SP2 machine with 2 GB of RAM and an Intel® Core™ 2 Duo CPU at 1.40 GHz. Consequently, the Windows version of the alternatives described above has been installed and run on the sample objects.

ImageMagick and GraphicksMagick are both in the Planets service registry [15] that is integrated into Plato. Experiments with both tools could thus be executed via web services, which facilitated experiment set-up, execution, and comparison of the original and the migrated file using JHove. Plato displays both JHove trees side by side, thereby allowing a thorough examination of file characteristics.

After having conducted the experiments we evaluated how well the migration tools preserved our defined requirements. This was done based on the requirements specified in the objective tree.

Results and conclusion

Based on the evaluation and possible outcome that is transformed to a uniform scale between 0.0 and 5.0 (0.0 means unacceptable) for each requirement, Plato gives a ranked list of the alternatives. This can be observed in Figure 2, which shows the aggregated performance values for each of the high-level objectives. Alternatives at the root level obtain the overall performance value and thus inform the planner which alternative meets the requirements best.

Figure 2 Utility values for high-level requirements

In this evaluation 'Keep status quo' excels over all other alternatives and is thus our recommended preservation action. The alternative, migrating from TIFF to JPEG 2000, was ruled out with GraphicsMagick, because the tool was not able to leave the image unaltered. The direct pixels comparison, using both GraphicksMagick's and ImageMagick's compare functionality, indicated that pixels had been changed during migration from TIFF to JPEG 2000. In the transformation we defined this requirement as a drop-out criterion, as the tool must be able to fulfill this requirement.

A detailed analysis of the result tree in Plato (partly displayed in Figure 2) reveals that none of the tools was able to sufficiently fulfill our object characteristics requirements (see Figure 1):

- All the applied tools show weak performance in preserving the ICC profile. The original ICC profile of the TIFF image was replaced by a default profile in the JPEG 2000 image.
- Besides the ICC-profile, the applied tools also didn't preserve the resolution of the scanned images. Using ImageMagick's and GraphicksMagick's identify command and XNView to analyse the migrated file's metadata indicated that the resolution as well as the TIFF properties were lost.

The same counts for technical characteristics.

- Migration to JPEG 2000 achieved lower values for 'Ubiquity' and 'OCR possible'. The OCR software used in the BSB can't handle JPEG 2000 files without migrating them to TIFF beforehand.
- At the moment, the advantage of using JPEG 2000 as master copies wouldn't spare us the effort of creating JPEG files for presentation, due to current insufficient browser support. Thus JPEG 2000 was not able to outperform TIFF in 'master = access copy', which was another technical requirement of our preservation plan.

Consequently, at this point in time not migrating the TIFF v6 images is the best alternative. However, in one year we'll look at this plan again to see if there are more tools available and whether or not the ones we considered in this year's evaluation have been improved. Furthermore, by then JPEG 2000 might be better supported by browsers. This would enable us to use the master copy as the access copy as well, which would reduce total memory consumption and the number of data files to be managed in the repository, since our 2 - 3 access copies would become obsolete. Both of these situations could change the recommended preservation action.

One option may be extracting the ICC profiles as well as the embedded metadata and storing them as a separate

file. Because this would mean a change in the established workflow and a significantly large change to the archival system, a detailed evaluation would need to be done. Once this evaluation has been done, we will revisit the current plan.

A cost evaluation of our preservation plan will certainly be necessary. Direct storage costs are decreasing, but detailed process and cost models regarding the complete life cycle of a digital object are still being worked on, and it is not possible to make precise statements about costs at this time.

All the trigger conditions mentioned above have been captured in Plato and are an important part of the preservation plan.

Designing this preservation plan showed very clearly that preservation planning is an indispensable function that every content-holding institution like BSB with a mandate for long-term preservation shall practice. Preservation planning is a process that should take into account the institution's individual requirements and policies, and should never simply be adopted from the plan of another institution. In fact, every institution individually must decide how to proceed within its particular environment.

Note

1) The objectives of the cooperative project between the BSB and the LRZ are the improvement of the organisational-technical infrastructure that has been developed as a pilot-system within the DFG-project "Long-Term preservation of Web Publications" with the aim of building a trustworthy and scalable digital archive as part of a national network for digital preservation, and the evaluation of the overall system BABS, the "System for Archiving and Access" (http://www.babs-muenchen.de). The project is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

References

[1] Paolo Buonora, and Franco Liberati. A Format for Digital Preservation of Images. A Study on JPEG 2000 File Robustness. *D-Lib Magazine*. Volume 14 Number 7/8, 2007. Available at <doi:10.1045/july2008-buonora>. (Accessed: November 5th, 2009.)

[2] Further information on JPEG 2000 may be found at <http://www.jpeg.org/jpeg2000/>.

[3] Robèrt Gillesse, Judith Rog, Astrid Verheusen. Alternative File Formats for Storing Master Images of Digitisation Projects, v. 2.0. Den Haag: Koninklijke Bibliotheek, March 2008. Available at: <http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/Alternative File Formats for Storing Masters 2 1.pdf>. (Accessed: November 5th, 2009.)

[4] *Planetarium*. The News Bulletin of the Planets Programme. Issue 7, July 2009. Available at: <http://www.planets-project.eu/docs/newsletters/Planetarium7_July09.pdf>. (Accessed: November 5th, 2009.)

[5] Bavarian State Library: <http://www.bsb-muenchen.de>.

[6] Stephan Strodl, Christoph Becker, Robert Neumayer, and Andreas Rauber. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, pages 29-38, Vancouver, British Columbia, Canada, June 2007, <doi:http://doi.acm.org/10.1145/1255175.1255181>.

[7] German Research Association: <http://www.dfg.de/en/index.html>.

[8] Munich Digitization Centre: <http://www.digital-collections.de>.

[9] Leibniz supercomputing centre: <http://www.lrz-muenchen.de>.

[10] English version of the Nestor criteria catalogue: <http://www.nbn-resolving.de/?urn:nbn:de:0008-2006060703>. (Accessed: November 5th, 2009.)

[11] Hans Hofman, Planets-PP subproject, Christoph Becker, Stephan Strodl, Hannes Kulovits, and Andreas Rauber. *Preservation plan template*. Technical report, The Planets project, 2008. Available at <http://www.ifs.tuwien.ac.at/dp/plato/docs/plan-template.pdf>.

[12] Christoph Becker, Hannes Kulovits , Andreas Rauber, and Hans Hofman. Plato: A service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL'08)*, pages 367-370, 2008, <doi:http://doi.acm.org/10.1145/1378889.1378954>.

[13] *ISO. Open archival information system – Reference model (ISO 14721:2003)*. International Standards Organization, 2003.

[14] For further information for the EU-project IMPACT (Improving access to text), see <http://www.impact-project.eu/>. (Accessed: November 5th, 2009.)

[15] Planets external deliverable PA3/D5: Report on Glossary and PA tool registry. <http://www.planets-project.eu/docs/reports/Planets_PA3-D5_Report_Glossary_Registry2.pdf>. (Accessed: November 5th, 2009.)

[16] The Wellcome Library: <http://wellcomelibrary.blogspot.com/2009/09/wellcome-library-to-use-jpeg2000-image.html>. (Accessed: November 5th, 2009.)

doi:10.1045/november2009-kulovits