callas software gmbh

EU Publications Office makes strategic decision to publish only in PDF/A format

The EU Publications Office converts 11 million paper pages to digital PDF/A documents for long-term archiving in its digital library

The EU Publications Office has chosen PDF/A as the format for its steadily growing digital library, currently comprising more than 150,000 documents. As the publisher for institutions of the European communities and the European Union, the office has the goal of making all contents available in digital form, including the numerous paper documents in the archive going back to 1952. This should make access easier for citizens and also ensure transparency.

The EU Publications Office had to develop a tight schedule plan to combine the paper-based and electronic archives with heterogeneous PDF files. Each day, the office, which is based in Luxembourg, publishes the official gazette of the European Union in up to 23 languages as well as materials for about 370 institutional author services. Then there are the online services, which provide citizens with access to information. This includes the "EU bookshop" service, which catalogs, inventories and archives the new publications each day as well as all historical publications of the EU. These are then available for download as PDF files.

As Dr. Silke Stapel, head of the EU Bookshop department of the EU Publications Office, explained,

"Citizens and companies should have access through a single facility to the publications of the European institutions, agencies and other resources. About 600 new publications are added each year to the rather extensive inventory."

Shortly after the EU Bookshop was set up, about 65,000 bibliographical notices were made available at (http://bookshop.europa.eu), and a scan-on-demand service was established in-house. Interested parties were able to use this to request PDF copies of the documents desired. However, the internal processing of this scanning service soon reached the capacity limits, and before long a PDF order could take several months to complete.

A uniform, complete digital archive for the EU bookshop

In order to meet the fast growing demand for PDFs and build a completely digital archive of all publications provided by the EU Bookshop, the decision was made in the fall of 2007 to carry out a digitization project. This required first of all that about 130,000 paper-based publications from 1952 to 2002 in eleven languages be digitized.

As project coordinator Anton Zagar from the EU Publications Office recalls, "Altogether this was about 11 million pages of heterogeneous content which was stored in the cellar archive of the main building." These had three requirements to meet for their digital future: reliable long-term archiving, the option of printing on demand and presentation that works well online in a web browser.

Anton Zagar, Project Coordinator EU Publications office

"For us, the decisive criteria were not only quality and the range of functions for the product, but also the outstanding support from callas software"

In the course of the eight week pilot project, 1.1 million pages were scanned, followed by the remaining 10 million pages between September 2008 and July 2009. A strict, detailed schedule ensured timely completion of the project. As Dr. Stapel explained, *"The project was to be completed in time for the Frankfurt Book Fair in October 2009. This deadline was not negotiable."* In two week cycles, the documents were picked up, scanned and converted to PDF/A files with pdfaPilot for secure, long-term availability. Final quality assurance was carried out by the EU Publications Office.

Conclusion

The project was successfully completed in November 2009 with the transfer of the final documentation. While the project was still ongoing at the end of 2008, PDF/A was also prescribed for the current production of publications. Moreover, since April 2009, the official gazette of the European Union has also been published in this format.

"This huge project could only be accomplished in such a short time thanks to the commitment of the staff and service providers involved. Many long nights and weekend work hours as well as close, nearly constant contact with the scanning service provider were determining factors for success," explained project coordinator Zagar. In conclusion, Dr. Stapel added, *"For projects of this scope, somewhat more preparation time is desirable. We also learned that quality requirements must be considered carefully in order to be clear about the technical, financial and manpower resource requirements as well as the timeline. Things usually take longer than planned, even when a time buffer has already been included."*

EU Publications Office makes strategic decision to publish only in PDF/A format

Anton Zagar and Dr. Silke Stapel

"The goal at the end of the project was to have a single file for each publication"

The format question

With regard to format, the conclusion was reached that the quality of all PDF files stored in the electronic archive should be standardized, including all future publications. Sharing experience with other libraries and archives as well the information from the conference held by the PDF/A Competence Center in the spring of 2008 in Amsterdam led to the choice of the ISO standard PDF/A-1b.

The specifications stipulated the production of TIFF 6.0 from scanning and subsequent embedding of the images in PDF/A-1b. Moreover, for print-on-demand capability, support of the ISO standard PDF/X-3 was required. Another important criterion was the provision of XMP metadata.

As Zagar explained, *"The goal at the end of the project was to have a single file for each publication, which contained the book cover and text, a searchable text layer, bookmarks, a thumbnail view of the cover and enough embedded metadata to provide a basic bibliographic description, good indexing and appropriate search results."*

More or less as a "byproduct" of this effort, the archive was to be inventoried and sorted out. As part of this effort, the duplicate copies which were always maintained there and had previously been kept next to each other on the shelf were to be moved to separate locations for security reasons, and publications for which no second copy had been made were to be identified and copied.

Outside specialist handles the large contract - and uses pdfaPilot

The digitization specialist Infotechnique SA in Luxembourg supported the project and scanned the paper archive. *"Our in-house scan-on-demand service did give us some internal scanning capacity and experience, but the success of this service had led to a considerable processing backlog. It often took months until a desired document could be provided. The enormous scope of the project and its uniqueness were also critical factors in the decision to outsource,"* said Dr. Stapel.

Together with Infotechnique, the EU Publications Office first carried out a pilot project in which various PDF/A tools were also tested. The pdfaPilot from callas software scored for performance and functionality, particularly because it was the only software to offer the required PDF/X support for printing on demand. Along with its validation capabilities for PDF/A documents, it also distinguished itself with extensive conversion options and comprehensive functions with which company-wide archiving strategies could be implemented according to the ISO standard.

Another strength of pdfaPilot is its XMP metadata editing, which was particularly important to the EU Publications Office. XMP stands for "Extensible Metadata Platform"; it is an open-ended platform for any custom metadata. The PDF/A standard requires all custom metadata to be organized in XMP schemas, which must also be embedded in the PDF/A file. Thus these are a prerequisite for the extensibility of metadata in PDF/A. The flexibility of XMP metadata achieved this way enables them to be used in any company- or sector-specific contexts. Apart from a program for embedding such custom XMP metadata structures, software support for creating and embedding a corresponding schema was thus also required. Here too, pdfaPilot was a winner.

"For us, the decisive criteria were not only quality and the range of functions for the product, but also the outstanding support from callas software", said Zagar, explaining the choice of pdfaPilot.