



Interim report for demonstrator – Germany
26th June 2004 to 30th June 2005

Contract No.:	<i>EDC- No. 11173 reUSE!</i>
Project full title	reUSE digital master files of printed material!

Deliverable number/name	<i>D-2.3 (according to Annex I)</i>
Security (distribution level)	<i>Public</i>
Contractual date of delivery	<i>T12 = June 2005</i>
Actual date of delivery	<i>11th July 2005</i>
Type	<i>Interim Report</i>
Status & version	<i>Final</i>
Number of pages	8
Author(s)	<i>Sven Bahnik, Matthias Schulz, Susanne Dobratz</i> <i>UBER</i>
Task responsible	<i>UBER</i>
Other contributors	

1	Introduction	3
2	Description of Work.....	3
2.1	Approach	3
2.1.1	Objectives.....	3
2.1.2	Focus Groups.....	3
2.2	Relationship with contents Providers	4
2.2.1	Address research	4
2.2.2	Contacts through e-mail, phone and meeting.....	4
2.3	Customer Relationship Management	4
2.4	Value Added Services	4
2.4.1	Print On Demand Service - ProPrint	4
2.4.2	HWO	5
2.4.3	Search Possibilities.....	5
2.5	Legal aspects	5
2.5.1	Copyright.....	5
2.5.2	Author agreement.....	5
2.5.3	Creative Common Licenses	5
2.6	Technical aspects.....	6
2.6.1	Submission	6
2.6.2	Internal Workflow	6
2.6.3	Repository	7
2.6.4	Data Volumes	7
2.7	Outlook.....	8
3	Conclusion.....	8
4	References	8

1 Introduction

The development of the edoc-Server starts with the project „Digital Dissertations“ in September 1997. In October of the same year the first dissertation was published on our system and on the end of the year the edoc-System offered the first Word-template and the first SGML-DTD.

The year 1998 started with the decision of the Academic Senate of the Humboldt-University to allow digital publications in order to fulfill publication duties.

In the third year in 1999 the first metadata database and workflow database was established.

The changeover to standard routines and standardised tools, formats and interfaces was realised in the following two years.

In 2002 the first server policy was published and we institutionalized the electronic publishing on the 2nd March of the same year.

Since October 2003 we use the XML-DTD xDiml and a conversion technology via StarOffice.

Since the beginning in the year 1997 we take part in several projects to advance the edoc-System. The objectives of the reUSE project are the set up of trusted digital repositories, the creation of added value services such as print-on-demand and the set up of an evaluation framework. It is the chance to follow our development road map.

2 Description of Work

2.1 Approach

2.1.1 Objectives

The Document and Publication Server offers the organisational and technical framework for publishing digital scientific documents. In course of this joint services offered by the Computer and Media Services and the University Library scientific documents of high importance are being published on the internet under strict quality control.

2.1.2 Focus Groups

Based on the collection mandate of the Humboldt University Library one of the target groups are all members of the Humboldt University. It refers to digitally born documents as well as digital versions of printed documents.

Also included are significant historic documents from the University Library and other institutions that are digitised due to terms of content, conservatory aspects, or the requirements of place-independent use.

Additionally the Humboldt University contacts institutions and organisations of the WISTA in Berlin. WISTA is a local network of twelve non-University Research Institutes, six natural science faculties of the Humboldt-University of Berlin and more than 370 high-tech companies.

2.2 Relationship with contents Providers

2.2.1 Address research

The faculties and institutes of the Humboldt University and the circa 370 high tech companies of the Science and Technology Campus Adlershof are the potential costumers of reUSE. We established an address list of this institutions and organisations.

2.2.2 Contacts through e-mail, phone and meeting

We invited all in the address list seized institutes to take part in the reUSE meeting in autumn of 2005. The meeting will take place in the Erwin Schroedinger centre and is organized in co-operation with the DFG project Scope. Topic of the one-day meeting will be the technical and organisational possibilities offered by the Computer and Media Service for authors and publishers and the Open Access initiative of the Humboldt-University.

2.3 Customer Relationship Management

A new category on the edoc webpage was created for the collection of open access publications. This new category contains publications of the nestor-project (<http://www.langzeitarchivierung.de>) and reports of the department of computer science. To collect and offer more open access publication an open access initiative at the Humboldt University was established. A special CMS-journal with the topic „Special issue of Open Access and edoc technologies“ was released to support the initiative.

It is important for content providers to have the possibility to influence the layout of their web presentations to get a more individual website. To satisfy this desire the layout of the edoc website was revised and supplemented with individually changeable elements. Now it is possible to individualize the website header picture and the used colours dependent on the type of document. With the change of the layout also the accessibility was increased.

The support of the authors and publishers takes place by E-Mail, at the telephone and in personal dialogue. To learn how to use the offered technical tools like the templates it is possible to visit one of our training courses or use learning material.

2.4 Value Added Services

2.4.1 Print On Demand Service - ProPrint

Tailor-made publications as paperback could be delivered from the Print-on-Demand-Service ProPrint. This service is based on the OAI-PMH interface. That is the reason why the implementation into an existing document- and publication server is pretty simple.

In the previous project period this Print-on-Demand-Service based on ProPrint was mounted on the online publication server TOBIAS-lib of the University library Tübingen. So ordering tailor-made publications as a book is possible. It is an enhancement of the library service.

We are still in contact with associates of the Technical University Chemnitz because they are also interested in adopting this service.

2.4.2 HWO

The project „HWO - HTWK Wissen Online“ at the HTWK-Leipzig is supported with technical and organisational solutions to establish a document- and publication server. Additionally several training courses with the topics electronic publishing and long term preservation were held in Leipzig.

This project revealed an additional aspect of Value added Service offering consultation and supply of basic technology like templates, database models and workflows for electronic documents.

2.4.3 Search Possibilities

We offer our user 3 possibilities to search for documents.

The first possibility is the search for metadata on our edoc-Server. This search used the OAI interface and is updated weekly.

We additionally offer two categories to search for online publications. In the first category are nearly 35,000 publications of 44 German archives searchable and on the second category are nearly 5,000,000 publications of 500 global archives searchable.

The last possibilities are the search via DDB with the pre-setting only search for online documents or via NDLTD.

It is planned to offer a full text search, but there are still a few problems to solve, so we can't offer it yet.

2.5 Legal aspects

2.5.1 Copyright

Observations of copyrights and rights of use for third parties are solely with the authors or the editors of the digital documents.

2.5.2 Author agreement

The author submits the non-exclusive publication rights to the university library, so the university library could offer a worldwide internet access to the publication.

The authors are still the owner of their copyrights and it's possible for them to offer rights for other utilisations.

The agreement defines a submission publication format.

2.5.3 Creative Common Licenses

Offering Creative Common licenses is still in process. The decision to offer CC licenses was made but the organisational and technical conditions must be created.

There are a few problems to solve. For example there isn't legal certainty about the question "What is the exact meaning of 'transforming'?" Is it transforming if you change the content of a document? For sure! But is it transforming if you only convert/transform a document from one format into another?

2.6 Technical aspects

This chapter describes the development of the edoc-System in the last 12 months, with attention to technical aspects.

The OAIS reference model was the starting point and basis for the changes and the development of the edoc-system. The main development was made in the workflow (see chapter „Internal Workflow“).

2.6.1 Submission

To receive well structured documents from our authors we offer templates for different editors (Word, Star/OpenOffice, LaTeX), checking tools and styles.

Most documents on the edoc-server were originally written using Microsoft Word. The authors have to use our word template (dissertation-97.dot; http://edoc.hu-berlin.de/e_autoren/vorlage.php) and we offer an additional checking tool (check.dot; http://edoc.hu-berlin.de/e_autoren/checkliste.php), a checklist and a training course. The checking tool reviews the use of the template.

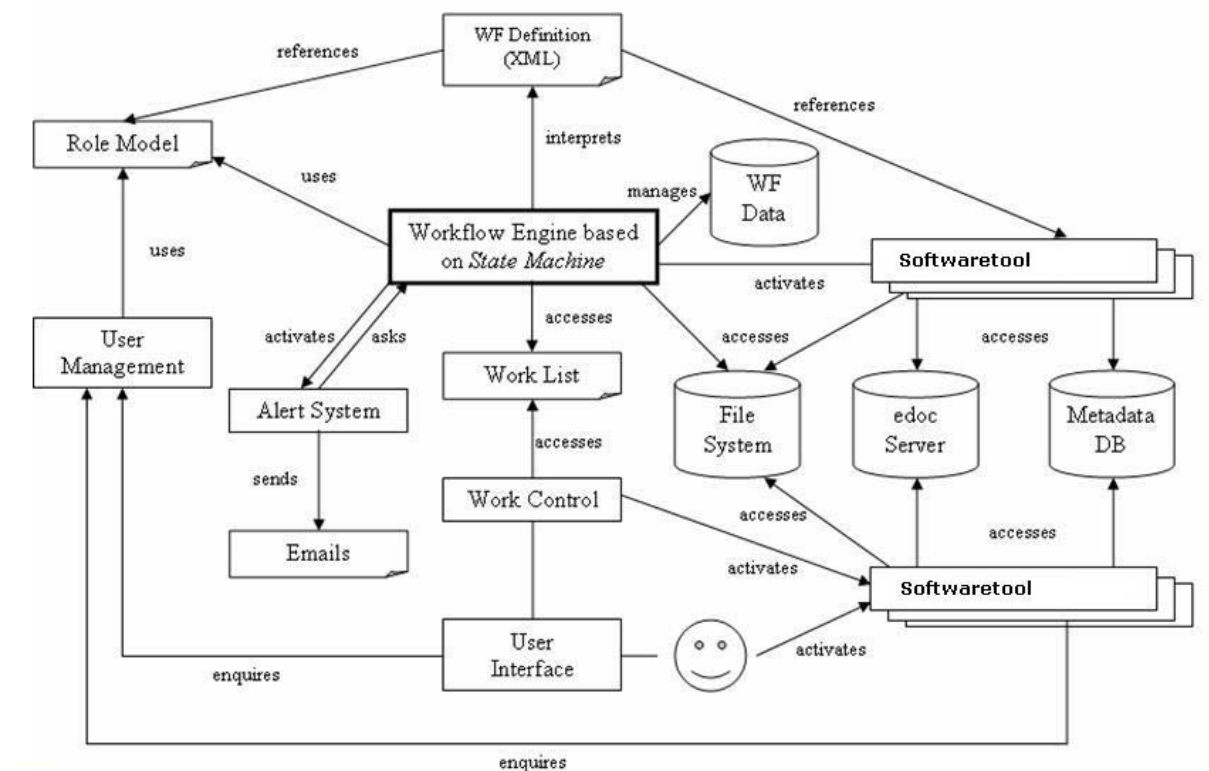
LaTeX users have to use a more flexible template, they can choose between 8 bibliography templates but they have to use BibTeX and we offer a positive and negative list.

PDF users can use a preflight-profile that checks a PDF-file for embedded fonts, Javascript use, security properties and the paper-format DIN A4.

2.6.2 Internal Workflow

The internal workflow is the centrepiece of the conversion and long term preservation process. The principal functionality of the workflow system is shown in the next graphic. This graphic is based on standard workflow ideas published by the Workflow Management Coalition.

Graphic 1: Workflow System



The workflow system itself is composed of a state machine, a role model and a work list. In addition it uses the file system and a relational database for storing its workflow data. Thereby, the workflow states can be permanently saved. Storing the workflow information other than in the main memory is especially essential to handle long running processes as they are common in publishing environments.

On the right side of the graphic you can find the external applications, the software tools (swt), that are controlled and whose actions are initiated by the workflow system. These software tools are responsible for atomic steps within a publication chain. In most cases these are functions to transform a document from one defined state to another. There are swts which are automatically activated by the workflow engine, and there are swts that are partly or completely controlled by human users of the system. The swts are integrated into the workflow system according to their formal specification. These applications may have access to the metadata database and the edoc-server or can handle user interactions.

The workflow engine is realized by an external state machine which is based on an open source project called Lucille, a project by the Austrian company XiCrypt. It is completely written in Java and allows full state and transition evaluation and control. As it is necessary in heterogeneous publishing environments the workflows can be easily modelled and altered using a XML based workflow definition language. This feature allows quick re-configuration of the workflow using standard XML editors. Additionally there exists a graphical editor for basic configuration. Using the workflow definition language, it is also possible to quickly incorporate new versions of swt which is necessary for the continuous development of these tools.

The state machine controls the flow and execution of the workflow. It supports the full spectrum of possible branching and parallelism. Every step in a workflow needs to be implemented using a specific java class. If the performed actions of several steps are alike, they can be implemented by the same Java class. Each step is afterwards handled by its own class instance. These classes can implement automatic action or a user interface necessary for human interaction. As mentioned above, many steps in the process of converting somehow need human interaction, due to the fact of manifold possible errors. We are continuously improving the tools, and some automatic error handling is already possible, but certain aspects as the correct layout of pictures and graphics cannot easily be decided by computers.

2.6.3 Repository

To manage the increasing amount of data (see chapter „Data Volumes“) and realise an improved service and handling for the content of the repository the edoc-server got a new hardware basis and the next step to a server that only store xml-files was made (decrease the number of different file formats). Also a new metadata database with an www interfaces was established and the changeover to dynamic XML and XSLT technology was made. The goal to have a XML- document management system including a search engine was reached.

2.6.4 Data Volumes

From June 2004 to June 2005 the number of documents in the repository increased from approximately 1650 documents to 2200 documents. In the same period of time the number of requests on the edoc-server increased from 2500 to 14700 per day.

2.7 Outlook

Non general work procedures up to the end of the project:

- implementing of Creative Common Licenses, expand the OAI interface with OAI-rightscomponent (rights of use could be propagated with the metadata)
- realization of further publication tools (templates, scripts, macros, conversion tools, styles, DTDs)
- market survey on promising technologies

3 Conclusion

The value added service ProPrint has been set up on the edoc-server and the University Library Tübingen, the technology and know how of the edoc-server has been shared with the HTWK Leipzig and the amount of collected and archived files still increase. The first steps to collect and archive documents from non-university institutions/organisations were made.

4 References

- edoc-search (http://edoc.hu-berlin.de/e_suche/)
- search via DDB (<http://opac.dbf.ddb.de:30080/DB=ext/MATSET?DBS=&TOP=O&YOP=&TAA=&LN D=>)
- search via ND LTD (<http://zipbo.vtls.com/cgi-bin/ndltd/chameleon?lng=en>)
- WISTA (<http://www.adlershof.de/index.php?id=112&L=1>)
- DFG (<http://www.dfg.de/en/index.html>)
- SCOPE (<http://edoc.hu-berlin.de/scope/>)
- Nestor (<http://www.langzeitarchivierung.de>)
- ProPrint (<http://www.proprint-service.de/>)
- TOBIAS-lib (<http://www.uni-tuebingen.de/ub/elib/tobias.htm>)
- University Library Tübingen (<http://www.ub.uni-tuebingen.de/>)
- Word template (dissertation-97.dot; http://edoc.hu-berlin.de/e_autoren/vorlage.php)
- Checking tool (check.dot; http://edoc.hu-berlin.de/e_autoren/checkliste.php)