Integrating Digital Document Acquisition into a University Library: A Case Study of Social and Organizational Challenges

Michael Hahsler
Department of Information Business, Vienna University of Economics and Business Administration
Augasse 2-6, A-1090 Wien, Austria
hahsler@ai.wu-wien.ac.at

Abstract: In this article we report on the effort of the university library of the Vienna University of Economics and Business Administration to integrate a digital library component for research documents authored at the university into the existing library infrastructure. Setting up a digital library has become a relatively easy task using the current data base technology and the components and tools freely available. However, to integrate such a digital library into existing library systems and to adapt existing document acquisition work-flows in the organization are non-trivial tasks. We use a research frame work to identify the key players in this change process and to analyze their incentive structures. Then we describe the light-weight integration approach employed by our university and show how it provides incentives to the key players and at the same time requires only minimal adaptation of the organization in terms of changing existing work-flows. Our experience suggests that this light-weight integration offers a cost efficient and low risk intermediate step towards switching to exclusive digital document acquisition.

Keywords: Digital Library; Integration; Document Acquisition Work-flows

1. Introduction

University libraries worldwide started to recognize that research habits significantly changed with the availability of research information on the Internet. Individuals, publishers, and other research organizations provide more and more research publications online. This makes the often time-consuming process of obtaining a printed copy through the local library faster and more convenient. Driven by the growing demand for digital document delivery university libraries spend a growing part of their budgets to license digital content from publishers and other information providers. In this setting it seems natural to think about making use of digital versions of research documents that are produced within the library's own university (e.g., theses and working papers). In the past librarians dealt with such documents in the same way as they were trained to deal with books and other documents that are obtained by the library in print. Since documents nowadays are almost exclusively produced in digital form (using word processing software) it would be a waste of resources not to make also use of a digital version.

Setting up a digital library has become a relatively easy task using the current data base technology and the components and tools freely available. However, to integrate such a digital library into existing library systems and to adapt existing document acquisition work-flows in the organization are non-trivial tasks. Several articles and project reports identify principles and guidelines for the successful development of digital libraries. For example, Harter (1997) identifies problem areas and formulates design principles in the

form of questions concerned with information quality, legal, and political problems. Rusbridge (1998) presents management lessons learned from the first two phases of the UK Electronic Libraries Programme (eLib) emphasizing the need for evaluation and continued training for project managers. Pinfield (2001) summarizes the lessons learned from the third phase of the same project including the areas of management, staffing, and partnerships. Wynne et al. (2001) present a practical guide for implementing a hybrid library consisting of 10 steps which include senior management support, team work, and strong links to the academic staff. McCray and Gallagher (2001) also present 10 principles for the development of a digital library including to design usable systems with open access, to be aware of data rights, to adopt standards, etc. From reviewing these articles it becomes clear that there exist certain guidelines but there is no simple solution to introduce and manage a digital library.

Virginia Tech was one of the first universities that started collecting electronic theses. In 1997 electronic submission was required by all students (Fox et al., 1997) and a purely electronic work-flow for submission, approval, and cataloging was established. This new work-flow completely replaced the former work-flow for printed theses. However, if a project to collect electronic theses is started in a university, changes such as requiring electronic submission, changes in the organization of how a school or dean approves a thesis, and changes how the thesis is cataloged by the library may be difficult to organize and coordinate. Especially in large universities such changes may be very slow since they affect many organizational units with very different requirements and sometimes rigid structures.

In this paper we present our experiences from the two year pilot phase of the project ePub^{WU} (electronic publications at the Vienna University of Economics and Business Administration) in light of the findings and principles identified in the literature. We do not focus on technical details but on the organizational and social aspects of the integration of a digital library into the university as an organization and especially into the university library. Not every organization can or is willing to change the work-flow of their theses instantly. For such organizations we show and discuss how a light-weight approach can be a successful first step towards acquiring electronic theses.

2. The ePub Project

In June 2001 the preparations for the ePub (electronic publications) project started with the formation of a project team consisting of a project leader and other personnel from the university library (from the recently created department of information and digital library) and developers from the university's academic department of Information Business. Senior management support was ensured by the university's Vice President for Research and the Director of the library. The aim of the project is to collect and manage the university's research-related documents (PhD theses, working papers, and later on master's theses) in electronic form and to make them available online. The necessary tasks (develop and implement a work-flow for the acquisition of the digital documents, provide information, and training for the authors and personnel, etc.) were assigned to the team members and a project schedule including the estimated cost were agreed upon. In October the project was officially launched and in January 2002 the Web site (http://epub.wu-wien.ac.at) went online including the first few PhD theses. In March the working paper section of the digital library was added and the archive was registered with the Open Archive Initiative (OAI, 2003). By June 2003 more than 300 scientific documents were available in the digital library and through the library's Online Public Access Catalog (OPAC). Currently, after a 2 year pilot phase, the results of the

project are reviewed by the library management and a decision on the future operation of the service, stronger integration into the document acquisition process of the library, extensions to other document types, and collaboration with other Austrian and international academic libraries will be made by early 2004.

3. The Information Life Cycle of PhD Theses

To organize this paper we adopt the research framework developed during the 1996 International Workshop of Social Aspects of Digital Libraries (Borgman et al., 1996). The framework covers three main perspectives of social aspects of digital libraries.

- **Human-centered view:** focuses on people, groups and communities. An issue is, for example, how to deal with heterogeneous user groups.
- **Artifact-centered view:** focuses on creating, organizing and representing information artifacts. Issues are making artifacts useful to multiple communities, integration of digital libraries into library catalogs (hybrid libraries), and professional practices of cataloging for digital artifacts.
- **Systems-centered view:** focuses on the digital library as a system that supports the interaction with digital artifacts. Issues are multiple interfaces for different user groups, open architectures for interoperability, and iterative development methods.

These views overlap considerably but all three views are necessary to cover the whole social problem space of digital libraries. Borgman et al. (1996) also developed a schematic information life cycle model to represent the information flow in terms of information artifacts and social processes. We adapted and extended the life cycle model in Figure 1 for PhD theses. The major phases in the model are characterized by the social context in which the information artifacts are used, namely: creation of information, searching for information, and utilization of information. In the creation phase we added the process of approving a thesis to the original model. In the original model this process can be seen as the very last part of the authoring process, however, for PhD theses the approval is of such importance that it is explicitly shown here as a separate process. To map the human-centered view onto the life cycle model we extended the original model in Figure 1 by the roles in which people and organizational units are involved in different phases of the life cycle.

For a successful introduction of a digital library it is important to provide incentives for cooperation to all involved groups. Providing incentives for authors is especially vital since this group is responsible for the authoring of the theses and therefore holds the rights to publish them. Although it seems largely beneficial for the authors to use the additional digital channel to make their work known there are considerable incentive problems. Fox et al. (1997) report from their efforts to collect digital theses at Virginia Tech that, even with a small financial incentive, only a very small proportion of their students chose to turn in their theses electronically. Therefore, the university decided to required electronic submission in 1997 which resulted in complaints by students due to the additional effort needed to prepare the electronic version. If requiring electronic submission is not possible or not sensible, finding incentives for authors is a key issue for the success of acquiring digital theses. Many PhD students are concerned that an electronic publication of their theses reduces their chances of publishing the results later in a high impact journal or as a book, but a survey of alumni of Virginia Tech revealed

that this is an unfounded worry (McMillian, 1999). Communicating this misperception effectively is vital for convincing PhD students to publish their theses online.

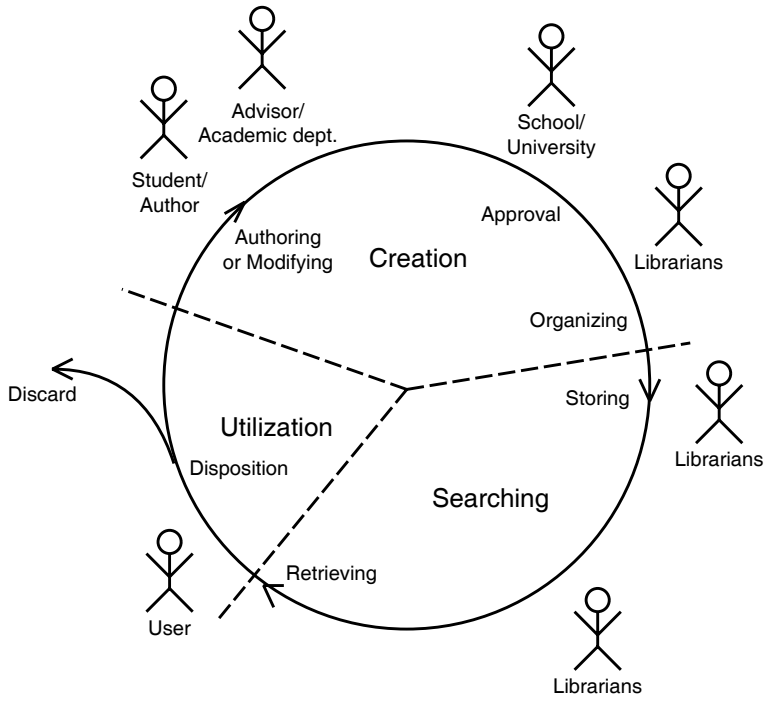

Figure 1. Information life cycle of a PhD thesis

The academic departments and the advisors who work with students have a massive influence over the students' decisions regarding their theses. Since students tend to follow the recommendations of their advisors (Wayne et al., 2001), it is important that this group has a positive attitude towards publishing online. Incentives for the advisors can be that online available theses demonstrate their ability to guide young researchers. For the department it can be used as a sign of the commitment to a high quality PhD program.

The school or university's main part in the information life cycle of PhD theses is that it sets standards and rules for the whole process. It sets standards on the format of the theses and governs and executes large parts of the work-flow (e.g., advising and approval). If the work-flow has to be changed for the introduction of a digital library, the university has to take the initiative. Many units of a university are affected by such changes which might be, especially in large organizations, a very slow and potentially political process that has to be planned and managed thoroughly.

The library is the obvious organizational unit in a university which can operate a digital library since the library's core business is to make information available to users. Integrating digital document acquisition and the management of a digital library offers the library the opportunity to gain experience with digital publishing and therefore opens new fields of activity and a range of strategic options (Geyer-Schulz et al., 2003). Inside a library exist many stakeholders with different perspectives. Library staff may develop or already have fears that the Internet with its ease to create and disseminate information in an uncontrolled way might undermine their quality standards and that information technology will cut down the need for trained librarians. However, it is more

likely that modern libraries will have to train their librarians and hire additional information professionals rather than substituting existing staff (Borgman, 2001). For the library management, next to strategic considerations, digital library services need to be economically sustainable. It is vital to analyze if the operation costs of the service are justified by the increase in service quality offered to the users, by the substitution of more expensive physical library services, or by a combination of both.

The library's users need to find and access information easily and in a cost and time effective way. Online availability of quality information is clearly what users more and more expect from their library. However, users of university libraries often do not see the massive cost associated with these services.

In Table 1 we summarize responsibilities, incentives, and problems for the identified roles of the information lifecycle of PhD theses.

Table 1. Responsibilities and incentives for people with different roles in a digital library for PhD theses

Role	Responsibilities	Incentives	Incentive Problems
Authors (PhD Students)	produce digital documents, agree to publish the thesis online	make work widely known for career advantage (e.g., to a prospective boss, the international research community, peers)	technical problems, additional effort and time, fear of disadvantages when publishing in a journal or a book based on the thesis
Academic Department and Advisors	influence the student to publish online	show his or her competence in guiding a young researcher, present the quality of research and of the PhD program	might publish advisors' ideas without giving sufficient credit
School/ University	support the digital library project, set standards for thesis submission and approval	show commitment to be an innovative and up-to-date organization, emphasize the quality of the PhD program	changes in standards and the work-flow of theses may be political, may take long and needs thorough management
Library and Librarians	integrate, operate and fund the digital library	improve services, gain experience in the online publishing process and investigate potential strategic options	additional work and cost, fear of losing importance due to information technology
Users	provide arguments for operating a digital library	convenient and easy way to find and access documents	users do not see the full costs of providing online documents

4. Adapting the Creation Phase for the Digital Acquisition of PhD Theses

In the creation phase the information is produced and then organized in an appropriate way. In Austria, as in most other countries, writing and submitting a PhD thesis is tightly regulated. There are legal restrictions as well as restrictions imposed by the university. At our university each PhD student has to write a thesis under supervision of an advisor. After the thesis is finished, she or he has to publicly defend it and submit the document to the dean of the school who assigns it to two professors for review and grading. After approval, the student has to deliver 2 copies of the thesis along with a form containing keywords, an abstract, and other bibliographic information to the university library where he or she gets a receipt. Only after the student finishes this process can he or she complete the PhD studies. As seen in this process, the university library only gets involved at a very late stage in the creation process (called organizing in Figure 1 above in this article).

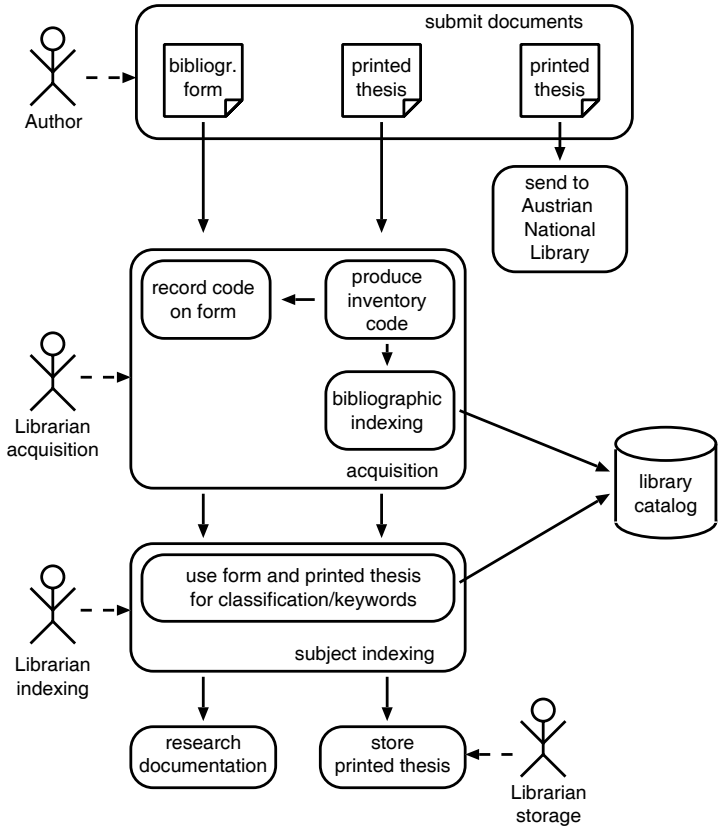

Figure 2. Existing work-flow for printed PhD theses

Figure 2 depicts the work-flow for the printed thesis inside the library. Several library departments have to collaborate for this process (denoted in Figure 2 by the roles acquisition, indexing, and storage). One copy is immediately sent to the Austrian National Library for archiving. The other copy remains at the university library and first undergoes the acquisition procedure. It gets an inventory code and the formal bibliographic data (author, title, etc.) is entered into the library catalog. In a second step a specialist classifies the thesis and assigns keywords based on the content of the

thesis and the form provided by the author. Finally, the printed thesis is physically stored at the library and the bibliographic form is used for research documentation.

To obtain a digital version of the thesis we first changed the form for bibliographic information filled out by the student into an electronic form that enters the information into the newly set up ePub system. This change did not affect any of the legal restrictions or school's policies and could therefore be done immediately. When the student fills out the online form he or she has also the opportunity to upload the thesis in a digital format. The student has to print the completed form and sign it to grant the right to use and pass on the bibliographic data (especially the abstract) to the library. If the student uploaded his thesis electronically this form also doubles as a contract which confers the revocable right to publish the thesis online to the library.

Since most bibliographic data was now available in electronic form we implemented a way to import the data from ePub into the library's catalog system, where it could be corrected by a librarian. However, this possibility was not adopted by the librarians who indicated that for the 200 theses a year it is more effort to correct the information given by the authors than to enter them from scratch using the printed thesis and the printed form. Moreover, the librarians involved in acquisition and subject indexing insisted that changes in the current work-flow are not possible because no personnel is available for additional tasks. Therefore, for the print version of the thesis, the whole process remained unchanged and the printed version of the filled out online form of bibliographic data is still used.

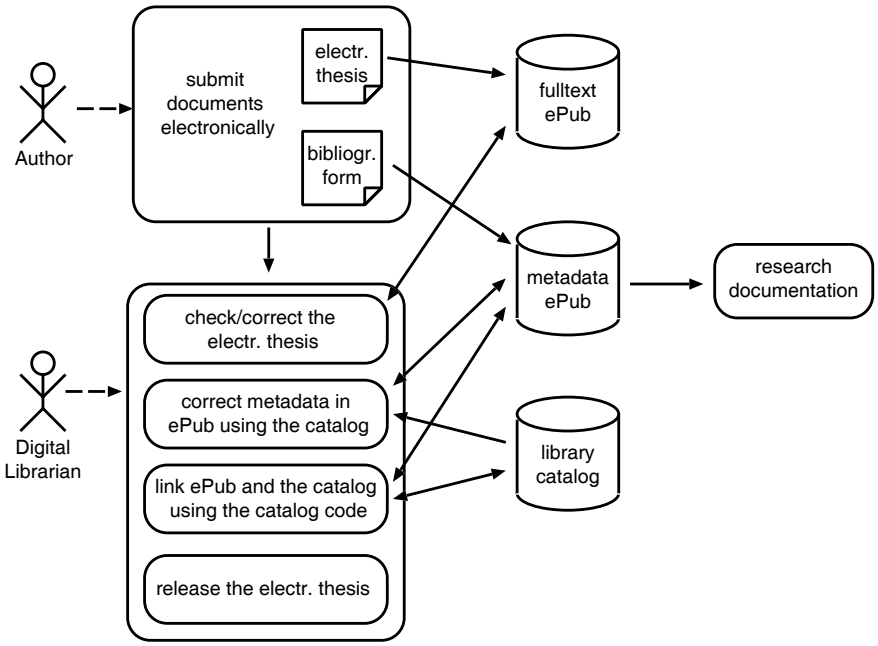

Figure 3: Additional work-flow for the electronic theses in ePub

For the electronic version of the theses a parallel work-flow (depicted in Figure 3) is carried out by a digital librarian from the library's department of information and digital library. This work-flow is started when a student who submits a printed thesis also provided an electronic version together with the bibliographic information. The digital librarian checks, corrects if necessary, and digitally signs the electronic version of the thesis. After acquisition and subject indexing is completed for the printed thesis, the

bibliographic metadata from the online bibliographic form filled out by the student is updated with the information from the library catalog which ensures that the system contains bibliographically correct information. A link between the library catalog and the ePub system is established by adding a reference (using the internal identification code of the catalog) to both systems. Finally, the electronic thesis is released and the electronic metadata inside the ePub system are used for research documentation.

5. Using the Searching and Utilization Phases to Provide Incentives

The searching phase in the life cycle consists of storing and retrieving documents. These two tasks mainly involve technical questions (e.g., retrieval technology, format conversion, and digital preservation) which are not discussed in this paper, but there are some organizational issues present. A big question is how the digital content can be accessed. For the library the Online Public Access Catalog (OPAC) is their main search tool and if a document is found in this catalog, which is also available in digital form, the user should have the choice either to download the document or to go to the library for a physical copy. In this view the digital library is seen only as a repository of digital documents and implements the exact functionality of its physical counterpart, the library storage. However, if only this functionality is provided, many important features like searching the fulltexts, providing departments with customized pages that contain online thesis supervised by their faculty, positioning of the documents in Internet search engines, and to network digital theses with other organizations would be neglected.

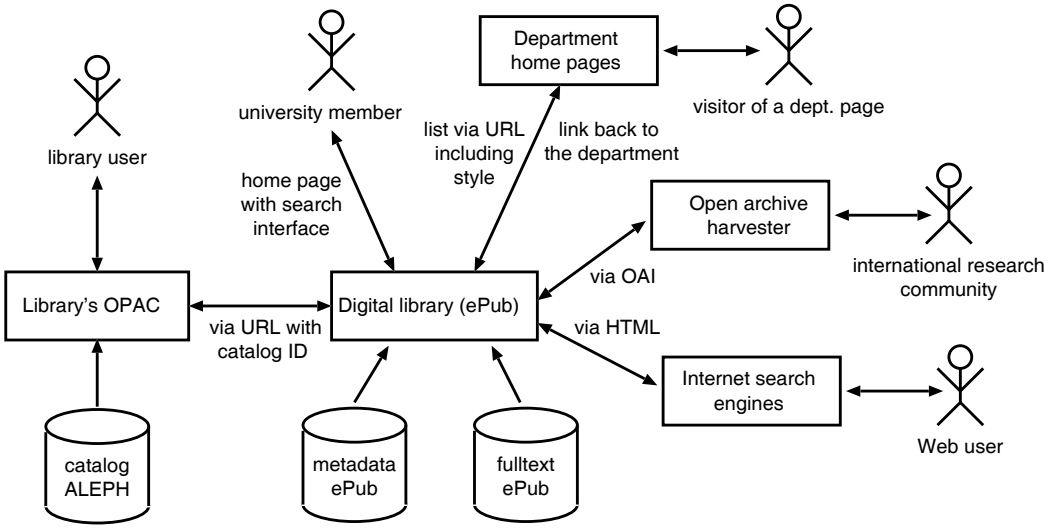

Figure 4. Multiple interfaces to support different user communities

Therefore, we implemented multiple interfaces, each supporting different user communities. Figure 4 depicts the different ways to access information in the ePub system. In the following we show how these interfaces provide incentives for the user groups identified above in this article.

- **Author:** For the author it is important that the online thesis helps for his or her future career and therefore can be easily found by prospective employers (e.g., in Internet search engines). For PhD students who want to pursue an academic career, it is

important that the thesis is present to the international research community (e.g., by providing an Open Archive Initiative compliant interface (OAI, 2003) for harvesters used by the respective research communities).

- **Advisor and department:** The department can link to the digital library for a list of theses written at the department. Such a list can include all features of the digital library and mimic the different styles of department home pages (see
- Figure 5). With this seamless integration a visitor of a department home page is not aware that the list is not supplied by the department but is a service of their digital library. Furthermore, every thesis in the library can be accompanied by a link to the department home page.
- **School/university:** The university can use the digital library to promote its PhD program and to integrate the digital documents into its research documentation.
- **Librarians and library:** Most library users search the library's online public access catalog. If a digital version of a found document is available in the digital library the user has the choice to download the digital document or to go to the library for the physical version (see

  Figure 6). This is the basic functionality that the library wants where still their catalog is in the center of the whole process. Book-keeping by the digital library is important for the library. Similarly to records of borrowed books, libraries need to quantify the usage of the digital library in terms of visitors and documents downloaded. This information can be used to estimate the value of the system by calculating the cost of storing and retrieving printed documents, producing and handling copies of documents under heavy demand, and by attributing cost to a out-of-stock situation which is quite common for libraries.

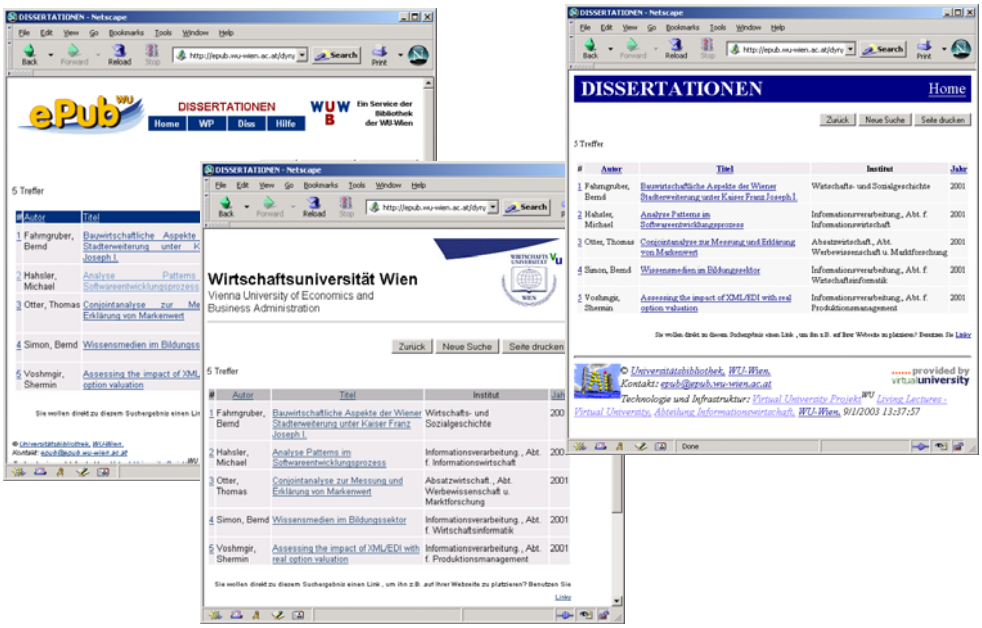

Figure 5. Mimicking different designs. From left to right: ePub standard design, official university design, design of the department of information business

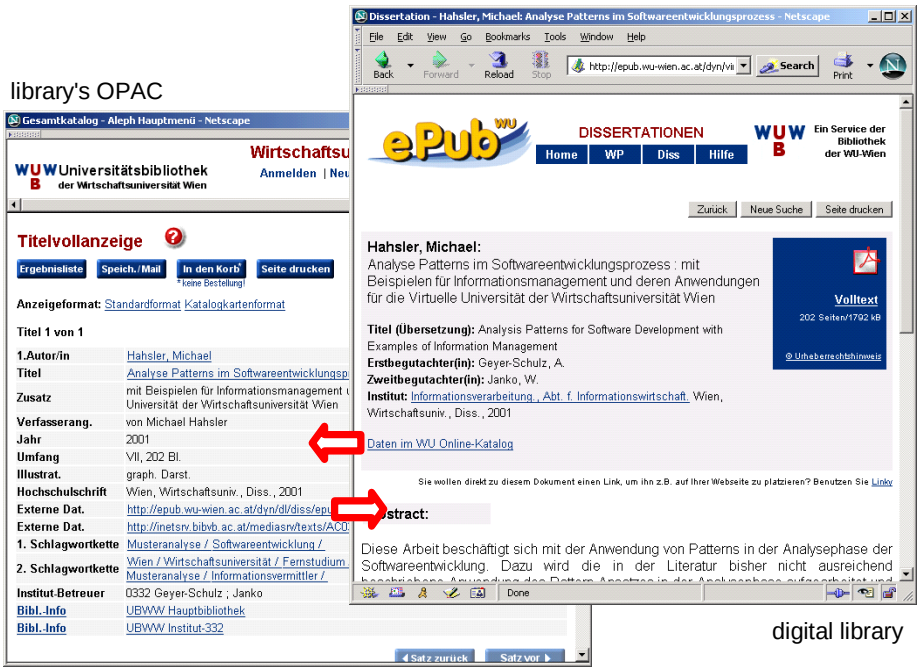

Figure 6. Bi-directional link between the library's catalog and ePub

6. Evaluation of the Light-Weight Integration of the ePub System

After the 2 year pilot phase it is time to evaluate the ePub system and the light-weight approach used for integration. The results are again organized following the phases of the information life cycle.

Creation phase: During the pilot phase we were only able to acquire 5% (10 theses) per year electronically from the authors. The main reasons for this low rate are exactly the same as already reported for other projects (e.g., Fox et al. (1997)). Most students said that they had already started a job and did not have the time to produce an electronic document suitable for the digital library. However, most of these students started writing their thesis before the start of the project and therefore did not get the information about the possibility of online publishing in time. Also with the availability of better conversion tools we expect this problem to decline. A second group, mostly teaching assistants from departments of the university, said that they did not want to publish their thesis electronically because they fear that this would reduce their chance to publish the results as a book. Several studies (e.g., McMillian (1999)) and examples from our experience show that this fear is unfounded but it is still hard to communicate this fact. Especially, since the current university guidelines state that for research evaluation of young researchers a published book counts as a publication while the online publication of the thesis by the library does not. Since resolving this incentive problem is a political question which takes time we added a clause to the contract between the author and the library which provides the author with the right to restrict the availability of his or her online thesis at any time to the campus of the university.

Since the work-flow for the printed thesis and the work for the librarians did not change for the light-weight integration reported here only little synergies are exploited. However, later in the project we also implemented a work-flow for the university's 8 working paper

series. The work-flow was radically changed and all working papers are now submitted to the library electronically using the ePub system. The acquisition and subject indexing is done together in the library's catalog and the ePub system. For the working papers the strict regulations and the incentive problems known from theses do not exist or are less important. By now, many working papers published before the project start and all new working papers (in total more than 300 papers) are already available in the ePub system.

Search phase: We analyzed the usage of the different interfaces of the ePub system for June 2003. The by far most important interface is the interface towards Internet search engines. Alone from the popular search engine Google we recorded 705 visits to the digital library. About half as many visits (319) came from various web pages within the university and 105 times a user from the university OPAC downloaded a digital document from the digital library. All users together produced 1016 queries for the ePub search interface and downloaded 560 documents. Due to the few documents in our system (and a high percentage of documents in German) only less than 10 visits resulted from the open archive interface in this month. This data shows that search engines are used today as the main interface to search for information and also research information on the Internet. Restricting a digital library only to be a document repository integrated in the library's catalog would mean to lose about 90% of usage and therefore potential value for the authors, the organization, and the users.

Utilization phase: For our library it takes up to one hour to obtain a physical copy of a thesis from library storage. Then the user can borrow the thesis for 4 weeks (with extensions up to 4 months). During this time other users interested in the same thesis have to wait. Therefore, it is impossible to provide access to the printed copy for more than a few users per month with only one printed copy of each thesis in stock. To lessen this out-of-stock problem for printed theses under high demand several copies would have to be produced at the cost of the library. Another cost factor is the risk that the only copy of a thesis is not returned. In this case the library has to request the copy from the National Library and produce a new copy.

For the electronic theses in the ePub system we recorded on average between 1.8 and 22.4 downloads per month (excluding automatic downloads by Web robots). This indicates that for theses under heavy demand only about 5% of the interested users would be able to get access to the one printed copy in the library or that the library would need to produce and handle about 10 physical copies at its own cost.

7. Conclusion

Overall, the pilot phase can be considered successful. The library explored its possibilities to take part in the digital publishing process and it built-up important knowledge in this field. More than 300 documents are already available online and these documents are accessed by more users than their physical counterparts were. The library management even already sees a potential of cutting cost by reducing the handling and stocking of theses and certain other research related publications if more and more of these documents are also acquired in digital form.

The light-weight integration of the digital work-flow of theses presented in this paper can be easily integrated more tightly with the physical work-flow or eventually replace the physical work-flow as already realized at Virginia Tech. However, its main characteristics are that it can be realized quickly since it does not require massive changes in work-flows and therefore is not slowed down by a bulk of political decisions

necessary in many parts of the organization. It is a first integration step that provides fast results to convince the authors and other stakeholders in the library and the university that a digital work-flow for theses and other research documents is feasible and in the end can offer advantages for all involved parties.

References

Borgman, C.L. (1996). Final report of the UCLA-NSF social aspects of digital libraries Workshop. University of California, Los Angeles.

Borgman, C.L. (2001). Where is the librarian in the digital library? Communications of the ACM , 44(5), 66-67.

Fox, E.A., Eaton, J.L., & McMillian, G., Kipp, N.A., McGonigle, T., Schweiker, W., & DeVane, B. (1997). Networked digital library of theses and dissertations. D-Lib Magazine, September 1997.

Geyer-Schulz, A., Neumann, A., Heitmann, A., & Stroborn, K. (2003). Strategic positioning options for scientific libraries in markets of scientific and technical information - the economic impact of digitization. Journal of Digital information, 4(2).

Harter, S.P. (1997). Scholarly communication and the digital library: problems and issues. Journal of Digital Information, 1(1).

McCray, A.T., & Gallagher, M.E. (2001). Principles for digital library development. Communications of the ACM, 44(5), 49-54.

McMillian, G. (1999). Perspectives on electronic theses and dissertations. Fourth International Conference on Grey Literature, Washington, DC.

OAI (2003). Open archives initiative. http://www.openarchives.org/

Pinfield, S. (2001). Beyond eLib: Lessons from Phase 3 of the Electronic Libraries Programme. http://www.ukoln.ac.uk/services/elib/papers/other/pinfield-elib/elibreport.html.

Rusbridge C. (1998). Towards the hybrid library. D-Lib Magazine, July/August 1998.

Wynne, P., Edwards, C., & Jackson, M. (2001). Hylife: Ten steps to success. Ariadne 27, March 2001.