The article provides an overview of the “Lehrstuhl” (professorship) for digital libraries at the Berlin School of Library and Information Science. The “Lehrstuhl” teaches both bachelor and masters students, and offers support for its doctoral students. The main areas of teaching are (1) research methods, with a particular focus on statistics and ethnography, and (2) digital libraries, with a focus on the development, scope, structure and metadata standards of digital libraries. Research covers the areas of long-term digital archiving – including “digital cultural migration”. The Berlin School is a partner of the LOCKSS (“Lots Of Copies Keep Stuff Safe”) long-term archiving network. A second research focus is on information behavior especially in the digital environment, and using ethnographic methods. Other research areas include the impact of copyright legislation and the history of information.

**Keywords:** Humboldt University Berlin, program of study, digital library, archiving, library and information science

---

**Introduction**

The goal of this article is to introduce readers to the “Lehrstuhl Digitale Bibliothek”, which was created in 2006. This is not a scholarly article, but rather an overview. Readers who want to look at some of the scholarly content should consult the publications listed in the references. In this article I will occasionally use the first person for cases when I am the sole or principle actor, otherwise I will use the term “Lehrstuhl” to describe collective projects. Since Lehrstuhl has no meaningful English equivalent, I will use the German term.
A digital library has two conflicting meanings. One refers to the independent digital collections that libraries, museums, and archives construct from their own collections. The other refers to the digital future of information organizations, including all cultural heritage institutions. While print on paper continues to fill the shelves of bookstores, scholarly information has become increasingly digital, and the use of academic journals has become almost entirely so, even if many libraries still collect paper copies. At the same time the sale of e-reading devices and electronic books and magazines intended for the general public has made significant progress toward general acceptance.

Libraries today cannot function without computers even to organize and locate paper materials, and as more and more content is offered exclusively in digital form, the transition to a truly digital library will become even more pronounced. The same is true for research data, which funding agencies recognize today as an essential part of scholarly publication and which libraries will need to accommodate in the future, if they are to remain content providers for researchers from all fields.

Teaching and research form the key components of any Lehrstuhl. The remainder of this article will address select elements from both.

**Teaching**

The following section will discuss curriculum elements where the Lehrstuhl has particular responsibility at the masters, bachelors, and doctoral level.

**Masters**

The Lehrstuhl is responsible for the two required courses in the standard MA program. Because these courses are the only required courses, they have disproportional influence on all students. It is important to remember that these courses evolve regularly. The goal is to keep students apprised of the latest issues and ideas, and not to fall into a routine of presenting materials that students could equally well read on their own.

The first of the required courses addresses research methods. The initial goal was to give students a broad overview of the range of methods that researchers at the Berlin School of Library and Information Science use for their scholarly work. This made the point that library and information science draws broadly from many disciplinary backgrounds, mainly in the social sciences and computing, but the approach provided too little focus and did not guarantee that students had the tools necessary to write a genuine contribution to scholarship as their master’s thesis. For this reason the course now focuses primarily on two topics.

The first is statistics, which has become as basic a component of scholarly education as reading and writing. The emphasis is not on deriving the mathematical proofs, but on understanding fundamental concepts like population, sampling, variables, and inference testing. These concepts often seem simple, but many studies fail because authors have not clearly identified their populations or have employed sampling techniques so flawed that their results are arguably meaningless. Students learn how to choose the right population, and how to define it so that it makes sense for statistical analysis. They learn also how to create a representative sample and how to test whether a convenience sample is sufficiently representative to be meaningful. Bayesian statistics receive attention along with conventional descriptive and inferential statistics. These are offered in a module on research methods, which will likely put even more emphasis on statistical analysis in coming years. A key point in the discussions about statistics is that these basic principles apply beyond purely quantitative studies and should be used in qualitative studies as well. Bayesian statistics received particular attention last year, in part because of their well-publicized and successful use in predicting the US Presidential election, but more particularly because Bayesian methods allow scholars to incorporate prior expectations from a wide range of studies that could not otherwise readily be incorporated into an analysis – a very typical situation in library and information science.

The second methodology to receive special emphasis is ethnography, in particular in the form of cultural anthropology as practiced by Clifford Geertz and his followers. The interest in using ethnographic methods to study libraries in particular began in the late 1990s in the US thanks to the work of Bonnie Nardi and Vicky O’Day (Nardi and O’Day, 1999) and has grown significantly in the ensuing years thanks especially to the work of Nancy Foster (Foster, 2007). Students are taught how to collect ethnographic data through observation and interviews, and they learn how to integrate survey data and quantitative information to develop a fuller understanding of particular user cultures, sub-cultures and micro-cultures. The ethnographic approach is as much as anything a way of seeing, hearing, and thinking, and it is well-suited to librarians and other professionals who must deal with the information needs of a broader so-
ciety. Human-computer-interaction can also be studied using ethnographic methods, but more specific approaches also come into play, including talk-aloud tests, extreme discount usability testing, and eye-tracking studies. Experimentation is also increasingly an element of HCI (human-computer interaction), in which log-files or other computer records are kept and analyzed statistically. Ethnographic methods have wide applications in other parts of information science, especially where marketing is involved.

Depending on the skills and interests of the students, other methodologies may also be discussed in the course, in particular computing-based techniques for information retrieval and legal issues having to do with copyright. One of the advantages of the Berlin School is that we can and do tailor courses to a certain extent to the strengths and interests of the students.

The second required course is on digital libraries. The focus of lecture portion of the course has long been getting students to develop their own definitions of digital libraries, instead of providing one for them. Digital libraries have changed substantially in their scope and purpose during the last decade. Students helped to develop a definition in a 2006 class that was later published as an article (Seadle, 2007), but this definition now seems too limited. The article is used nonetheless both as an example of how students can think about the issue and as an example of change. The course looks historically at the development of the concept of digital libraries, which involves the development of computing capabilities and network infrastructure. It is important that the students understand, for example, that the Internet existed long before the World Wide Web, and that our current Internet Protocol was not the only network option for providing access to content, but rather an option among many that survived in the international marketplace. The final weeks of the course generally look beyond traditional collection-based digital libraries to content providers like Europeana, Google Books, and even Amazon.

The seminar that accompanies the lecture has focused on the semantic web, in particular the capabilities of the Resource Description Framework (RDF), and on other standard metadata structures that underlie the access to networked information. It is important for students to understand the role of metadata in the computing infrastructure, especially at a time when the number of metadata standards has grown to the point that few actual common standards exist.

The Lehrstuhl is also actively involved in teaching in the Postgraduales Fernstudium, which may best be translated into English as “executive education”, since students are expected to have positions and/or experience and are generally seeking higher level positions. A digital library course is part of the curriculum and follows a similar pattern to the course described above. Another key course is “what is library and information science”, which explores how the field defines itself and uses a quantitative source, namely the iConferences, in the final session for determining who and what defines the field internationally. User studies and human-computer interaction are also taught, with an emphasis on ethnographic methods.

**Bachelors**

The Lehrstuhl offers regular courses at the bachelor’s level. (See, Umlauf, 2013 in this issue) This article will discuss only two. The first is called Information and Society and addresses the issue from the broadest possible perspective. Among the topics is an overview of library and information science organizations in Germany and other countries, in particular the US and often France or the UK, depending on the lecturer and that person’s expertise. The goal is for students to understand the range of organizations and acronyms they will encounter in their student as well as in their professional life, and to understand how very differently these organizations are structured in other countries. The American Library Association is, for example, primarily an individual-membership organization with a federal structure that includes divisions such as the Library Information Technology Association, and the Association of Research Libraries is open only to the top research libraries by invitation and election, but covers Canadian in addition to US-American libraries.

The Lehrstuhl also offers an introductory required course called Grundlagen der Informations- und Kommunikationstechnologie, which introduces students to the basics of computer programming in either Perl or Java. It is important that students can write simple programs to extract information from digital sources, and both languages provide a shell for students to use Regular Expressions, which gives them a standard and highly flexible tool for identifying complex word and letter patterns within a machine-readable corpus. These programming skills feed into the teaching and research of other groups.

---

1 [http://www.w3.org/RDF/]
at the Berlin School, in particular Information Retrieval and Knowledge Management. Many students come to library and information science programs with preconceptions that they will primarily be reading and writing, and they must learn how intensely modern libraries and other information services depend on computing techniques and computer-based services. Students who know at least the basics of programming have a better preparation for their future jobs.

**Doctorate**

Teaching at the doctoral level traditionally plays a minimal role at German universities, when courses are offered at all. The Humboldt-Universität zu Berlin has established a Graduate School that offers a wide range of training and support for doctoral students. The Lehrstuhl also offers a research colloquium together with the Lehrstuhl for Information Retrieval and the Royal School of Library and Information Science in Copenhagen. The twin goals of the colloquium are to discuss research methods and offer training in them, and to give doctoral students a forum where they can get feedback about their projects and their drafts from other doctoral students as well as from professors who may not be on their committee.

The Lehrstuhl does not formally limit the research methodologies and approaches that doctoral students may take, but as a practical matter there is a strong emphasis on ethnographic methods and on statistical tools. Those using ethnographic methods are generally pursuing a qualitative method and need to focus on interview techniques and to think about how to analyze the results, which may mean classifying transcripts, but that is only one of many approaches. Those engaging in experiments using either methods out of psychology or behavioral economics are doing fundamentally quantitative research and need more intensive statistical training, particularly an understanding of how they can use inferential statistics to answer research questions. Some of this training is available in other units of the university, but the research colloquium gives an opportunity to shape the training specifically to the needs of students doing library and information science research, and allows an individual focus on the dissertation projects.

American style doctoral programs regularly offer these kinds of colloquia and special programs, and more German universities are offering extended support for their doctoral students, in part because interdisciplinary research is growing and professors can no longer expect that doctoral students will have had all of the research training they need during their bachelors or masters programs, especially since these programs tend to be shorter and more intensely focused than the old Magister. While the colloquium offers advantages for doctoral students who reside in or near Berlin and can come regularly, the more traditional German doctoral student who has a job and may live in another city cannot readily profit from it. Videoconference could be an option for future participation, as well as more distance education content in courseware management systems like Moodle.

**Research**

The following sections will discuss a series of research projects and topics involving the Lehrstuhl and its doctoral students.

**Digital Archiving**

Long term digital archiving involves a broad range of topics, some of which are more library focused, such as collection strategies, and others are more computer science and engineering focused because they involve storage methods and integrity checking. The Lehrstuhl research involves both. Much of the discourse in the field is based on marketing claims from vendors and from outdated concerns that grow out of bad personal experiences with backups. The Lehrstuhl research involves both. Much of the discourse in the field is based on marketing claims from vendors and from outdated concerns that grow out of bad personal experiences with backups. The Lehrstuhl research aims at greater transparency in the claims and in the discourse about archiving generally, with a particular emphasis on the need for more publication of research results in peer-reviewed journals. An often repeated claim is, for example, that bit-preservation is easy. In fact a simple mathematical analysis of bit rot can show that the problem is very real for data held over long periods of time (See Rosenthal, 2005). Much concern about format change exists especially among librarians, but there is in fact little evidence of format-obsolescence for publication formats on the Web. (See Rosenthal 2010).

Two issues of concern for archiving are whether publishers favor one archiving system over another. In or-

---

2 http://www.iva.dk/english/

3 http://moodle.hu-berlin.de/
order to move the discourse out of the realm of vendor claims, I did a numerical study using data from the websites of Portico, LOCKSS (Lots of Copies Keep Stuff Safe from Stanford University) and CLOCKSS (Controlled LOCKSS). Since LOCKSS and CLOCKSS are just different business models for the same software, their results were pooled.

“The details of these statistics are by no means completely reliable. They rely on matches based on assumptions that the entries in the tables for over 25,637 titles are so accurate that the matches performed against them are reliable. It is important to remember that the processing understates duplication. Nonetheless a few striking results deserve comment.

The first of these results is the degree to which publishers, especially large publishers, are ready to work with multiple archiving systems. ... Smaller publishers are in a different situation. For them the membership costs in multiple archiving systems are sums as significant as for libraries – and for the very small publishers even more problematic. Very small publishers also have no scholarly or engineering basis for choosing one or another system. The preference for LOCKSS should be seen in significant measure as a financial decision.” (Seadle, 2011)

Authenticity is another complex archiving topic where the definition is far from precise and the measures in the digital world vary in significant ways from those for analog works.

“Authenticity in the physical world implies genuineness, and with physical objects a sense that it is the actual original, but the concepts of authentic, genuine, and original grow less clear the more closely an object is examined. Is, for example, a contemporary printed copy of Charles Dickens’ novel “Oliver Twist” authentic? Likely it contains many of the original words, but some words and phrases from the initial publication were corrected in later editions. ... Even in the physical world the concept of authenticity rapidly becomes open-ended once it is divorced from a specific object: this work in this version at this time.” (Seadle, 2012)

Digital authenticity is of course not about specific objects, but about copies and versions. One of the key differences is that a trusted repository plays a key role in analog authenticity, but trust plays no role or even a negative role in digital authenticity, since any digital work may be rendered inauthentic by internal or external attack. At the Vatican Workshop on “Long-Term Preservation... From the Stars?”, I argued that:

> “Today we measure authenticity chiefly with:
>  - check-sums or hash-values that show that copies are identical
>  - log file provenance records

> These measures are intolerant of the changes (e.g. aging, restoration or editing) that are acceptable as part of analog measures of authenticity.” (Seadle, 2012)

The relative intolerance of the measures is potentially a problem, because the working concept of authenticity has considerable flexibility. Nonetheless tying authenticity to a strict definition of digital integrity has the virtue of bringing it closer to notions about the integrity of physical artifacts, though even with original artifacts the toleration for change (restoration in paintings, for example) is far greater than these digital measures allow.

**Digital Cultural Migration**

The idea of cultural migration is new to digital archiving. It is an ethnographic concept that extends the idea of format migration. The goal of format migration is to make sure that future generations can open files in 100 or more years. The goal of digital cultural migration is to help future readers and users understand the content and its layers of meanings. The problem of cultural migration is easy to understand by looking at the past. Over time words can change their denotation or connotation. Today publishers automatically do a form of cultural migration by providing notes in older books to explain the language and to explain references to places or political events that readers may not find familiar. Not all works suffer equally from the need for footnotes or explanations, of course. Most of Jane Austen’s novels require little explanation today. The problem is not limited to text. Games, movies, pictures, even music can grow to seem old-fashioned and visual or musical references can become obscure.

Recognizing the need for digital cultural migration is arguably one of the jobs of a digital curator with appropriate ethnographic and technical training. While format migration is comparatively easy to recognize when a file is no longer readable, the need for cultural migration is harder to identify because it involves a complex of cultural factors. It may well be possible to write software to discover potential problems with names, places and well-known word changes, but other kind of cultural variance are harder to identify, even for humans, humor for example. Not all cultural phenomena can reasonably be migrated any more than all words are translatable. Nonetheless factual gaps can be solved with links to informa-

---

4 http://www.portico.org/digital-preservation/
5 http://www.lockss.org/
6 http://www.clockss.org/clockss/Home
The Lehrstuhl played a leading role in the Deutsche Forschungsgemeinschaft-funded LuKII (LOCKSS und KOPAL: Infrastruktur und Interoperabilität) project. This project proposes interoperability between LOCKSS and KOPAL in order to combine cost-effective bitstream preservation with an established tool for usability maintenance and format migration.

Based on these goals the chief elements of this project are: 1) to establish a cost-effective LOCKSS network in Germany including infrastructure to provide ongoing technical support and management for LOCKSS and its variants (e.g. CLOCKSS); 2) to conceptualize and implement interoperability between LOCKSS and KOPAL in order to combine cost-effective bitstream preservation with well-developed usability preservation tools; and 3) to test the interoperability prototype by archiving data from German institutional repositories. (Seadle, 2009)

The project is in its final phases with most of its major goals completed. In working on the third goal of harvesting open access repositories, the LuKII team became aware of legal issues that affected the scale of the test.

**Law and Information**

Legal issues have always been a factor for libraries and other information and cultural organizations, but the issues have become more urgent in a world where digital information is available ubiquitously. Before coming to Berlin I wrote a column on copyright issues in US law from the perspective of a user and ethnographic observer, and explicitly not from the viewpoint of a lawyer. Library and information science research may affect how laws are written, but the primary goal of my research has been to examine how people understand and interpret the rules. The rules for the use of protected materials are comparatively clear and strict in Germany. The rules for “fair use” in the US and for “fair dealing” in UK law allow far greater flexibility, and these laws also matter for German scholars, since significant amounts of scholarly information is digitally available in the Anglo-American domain. Another area of international significance is the set of rules for handling orphan works, that is, works that are presumably still under copyright protection, but for whom no rights holder can be found. This is particularly a problem with Web-based materials, where the true author may not be listed or may be hard to detect. Differing national rules about work-for-hire (in the US) and moral rights bring additional complications.

An area of particular interest is the technology used to detect putative copyright violations, among them “honey pots” that appear to offer free copyright-protected contents, but are in fact set up to capture the IP address of people downloading them. Young people with a limited understanding of the legal issues involved especially fall into the trap. While these honey pots catch consumers, there is little evidence that they hinder major abusers. The technology of detection applies not only to companies trying to protect their intellectual assets, but also to plagiarism detection, which has become a broad and politically sensitive topic. A number of software companies offer tools to detect copying, but even the best of them have weaknesses, especially in technical areas where ways of expressing research results may be highly standardized. The systems often work better for English than for German, because they have more English content for comparison.

---

LuKII

Legal issues have always been a factor for libraries and other information and cultural organizations, but the issues have become more urgent in a world where digital information is available ubiquitously. Before coming to Berlin I wrote a column on copyright issues in US law from the perspective of a user and ethnographic observer, and explicitly not from the viewpoint of a lawyer. Library and information science research may affect how laws are written, but the primary goal of my research has been to examine how people understand and interpret the rules. The rules for the use of protected materials are comparatively clear and strict in Germany. The rules for “fair use” in the US and for “fair dealing” in UK law allow far greater flexibility, and these laws also matter for German scholars, since significant amounts of scholarly information is digitally available in the Anglo-American domain. Another area of international significance is the set of rules for handling orphan works, that is, works that are presumably still under copyright protection, but for whom no rights holder can be found. This is particularly a problem with Web-based materials, where the true author may not be listed or may be hard to detect. Differing national rules about work-for-hire (in the US) and moral rights bring additional complications.

An area of particular interest is the technology used to detect putative copyright violations, among them “honey pots” that appear to offer free copyright-protected contents, but are in fact set up to capture the IP address of people downloading them. Young people with a limited understanding of the legal issues involved especially fall into the trap. While these honey pots catch consumers, there is little evidence that they hinder major abusers. The technology of detection applies not only to companies trying to protect their intellectual assets, but also to plagiarism detection, which has become a broad and politically sensitive topic. A number of software companies offer tools to detect copying, but even the best of them have weaknesses, especially in technical areas where ways of expressing research results may be highly standardized. The systems often work better for English than for German, because they have more English content for comparison.

---

7 http://www.lukii.hu-berlin.de/
Information History

Library history is an old and well-established field, but information history has only started to attract attention as information specialists began to look at their own past. Elke Greifeneder and I assembled an issue of Library Hi Tech (volume 30, number 4) that reviewed the last thirty years, during which many contemporary Internet resources were established. Gabriele Metzler of the History Department at Humboldt-Universität and I are collaborating on research program that assembles interviews and other documentation, including project proposals, to build a resource that will serve future researchers. Part of the plan is to link names and places automatically with references using the Resource Description Framework (RDF) to facilitate discovery. The range of information that should ideally be collected is broad, and includes not only existing versions of digital artifacts, but earlier versions, log file data about the usage, and user reactions. Some information has already been lost, and some judgment about what is ultimately necessary must be exercised because of the cost of collecting rather than the cost of storage.

The history of the digital present is a history of choices, many of which were far from obvious at the time and some of which may ultimately have been less than ideal. Its development can only be understood in the context of the time, the tools available, the ideas in people’s minds, and the resources available. The cold war, for example, influenced the development of the internet, which began as a military research project that needed to be robust during a nuclear war. Fears and misconceptions have equally influenced developments, for example bad personal experiences with rapid changes in early file formats influenced priorities for long term digital archiving. The machine-intelligence models of the 1990s and their failure to deliver on expectations has convinced many researchers that machines will never be able to comprehend natural language meaningfully, even though the statistical and database models used by Google and in the IBM Watson vs. Jeopardy contest give promise for new developments.

Information Behavior

The research of the Lehrstuhl is ultimately about how people interact with information. This is especially true for the legal and historical research, and plays a major role in digital cultural migration. One of the best examples of cutting edge research in the Lehrstuhl comes from the dissertation of a former staff member, Elke Greifeneder, now a professor at the Royal School of Library and Information Science in Copenhagen. Greifeneder (2012) wrote:

“User studies in digital libraries face two fundamental challenges. The first is the necessity of running more user studies in an online environment. Users can access digital library collections and services worldwide and the services should be usable at any time, because users access these services at a time and place of their choice. Online studies enable researchers to be separated from their participants in space (synchronous tests) and/or in time (asynchronous tests). This need for more online studies is coupled with a second need, a demand to test under realistic conditions outside of laboratories in users’ natural environment.”

This is fundamental research into how user research can reliably be carried out in online environments, and was awarded the Verein zur Förderung der Informationswissenschaft dissertation prize in 2013.

A great deal of statistically invalid research is done on online environments. There is a reasonable argument to be made that any results are better than none, but the work of Greifeneder argues strongly in favor of more rigorous statistical models and testing that takes the test-environment into account.

Conclusion

Digital libraries are no longer the logical focal point for research that they seemed to be in 2006, partly because they have grown well beyond their original scope and focus. Digital libraries arguably today encompass everything in the networked world. Nonetheless the focus of research in the Lehrstuhl has remained consistently focused on the user. Technology obviously plays a significant role, but focusing research on one limited aspect of how information technology facilitates developments threatens to narrow the scholarly focus to an application-level. For this Lehrstuhl the technology remains plastic and humans remain in the forefront of the inquiry.

Notes

[1] Most of this text comes from materials I prepared for students studying long term digital archiving at the Berlin School of Library and Information Science.

8 http://www.vfi-online.org/
References


Prof. Michael Seadle, Ph.D.
Humboldt-Universität zu Berlin
Institut für Bibliotheks- und Informationswissenschaft
Dorotheenstraße 26
10117 Berlin
seadle@ibi.hu-berlin.de
www.ibi.hu-berlin.de