# UNIVERSITY LIBRARIES – BETWEEN SERVICE PROVIDERS AND RESEARCH INSTITUTIONS

WOLF-TILO BALKE

Institute for Information Systems (IfIS), TU Braunschweig, Germany
balke@ifis.cs.tu-bs.de

**Abstract**

In the last years the process of generating, disseminating, and archiving new knowledge has changed fundamentally. Beside the increasing amount of new knowledge that needs to be processed, new paradigms for search, access, and exchange have evolved: digital information is discovered, interlinked with curated databases, commented upon, adapted, and shared in Web-based collaborative research infrastructures. And this does not only concern classic scientific publications in monographs, journals, or conference proceedings, but also data in the form of models and simulations, experimental data sets or results of analyses. This new way of creating knowledge is often referred to as e-Science (enhanced science), and heavily relies on modern Web information management and Web 2.0 technologies.

In order to reflect these changes university libraries as dedicated infrastructure provider for the management of scientific research information need to get active: besides handling the exponentially growing amount of rather heterogeneous material (the information deluge) users need customizable and personalizable digital tools and value-added services to support the effective and efficient utilization of information resources But the key to providing a solid foundation for all such services obviously lies in the comprehensive and qualitatively sound indexing of the textual and non-textual resources, which, however, usually needs manual efforts and thus is hard to provide on today's limited budgets. In this paper we will take a closer look at the current change in libraries and the research challenges to develop advanced methods for information provisioning. We argue that this research needs to become an integral part of modern library structures and will be strongly interdisciplinary in nature.

**Keywords:** digital libraries, metadata generation, proactive services, personalization

## Introduction

The way in which output of academic research is produced today has considerably changed in the digital age. On one hand the sheer amount of information produced by academic research and published in scholarly outlets grows continuously. The stm report 2015 states that for the last 200 years both, the number of articles and the number of scholarly journals grows by about 3% per year, though there seem to be some indications that growth even has accelerated in recent years. This constant growth appears to be mainly driven by an equally persistent growth in the number of researchers of about 3% per year (Ware & Mabe, 2015). On the other hand, the content offered by university libraries tends to grow much more heterogeneous: from mostly textual content in books, conference proceedings, and journals to all kinds of digital content like audio-visual content, primary research data, and complex simulations or models.

While for university libraries this already has severe implications in the areas of (semantic) indexing and interfaces for content provisioning, also the usage of resources has changed. As a general trend search behavior gets more complex and often has to involve different sources like literature, curated databases, and experimental data, i.e. the traditional index search becomes a sophisticated search process that often needs different support tools in different stages. This is of course due to the growing complexity and specificity in today's scientific domains, but also due to an often interdisciplinary research methodology needing to use several conceptualizations (like thesauri, taxonomies, or ontologies) within the

same search process. Because there is also an increased need for the individual library user's efficiency, this leads to the development of specialized digital services: the traditional library search interface becomes an intelligent discovery system.

This need has also been recognized by prominent funding agencies like the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), which recently phased out one of its oldest grant programs for libraries, the system of *Special Subject Collections*, and introduced a new funding line called *Scientific Information Services* (Kümmel & Reinhardt, 2011). The idea of the old program mainly was to distribute responsibilities for building up possibly complete literature collections for each discipline. This basically means a balancing of funds needed for acquisitions over the German universities' library system to avoid duplicate acquisitions of necessary, yet rarely used items in a field. The idea of the new funding line is to additionally provide (pro-)active support for knowledge acquisition in each discipline: a new challenge for almost all university libraries. The decision for this radical change in funding guidelines followed a thorough evaluation of the old system using DFG proposal-processing data and intensive interviews conducted with officials from 15 selected libraries.

In this position paper we argue the research and development of innovative and value-adding services needs to become an integral part of today's library structures. This includes the new challenges of traditional issues like for instance content indexing, where a high-quality semantic indexing of different content types for a variety of purposes has to be performed. In times of rapidly growing amounts of content and usually rather limited funds this indexing can no longer be performed manually by dedicated library departments, but has to involve some degree of automatization and/or contract an (voluntary) external workforce, e.g. by crowd-sourcing tasks or monitoring interaction in the Web 2.0. In any case, rigid measures for quality control are needed. But what is more, some new challenges like the development of discovery services will not only need a thorough understanding of current research, but also an intuition about future directions in different disciplines paired with expertise in user experience and new business models.

**Content Enrichment by Text Mining and Automatic Metadata Extraction**

The most important step in making scientific results accessible to the public is a proper indexing to allow for effective retrieval functionality. But in the current information deluge, where more and more scientific objects of rather heterogeneous types (publications, primary research data, or audio-visual content) and from rather specialized disciplines need to be handled, deriving indexing terms in sufficiently high quality is getting harder: while funds for manual indexing by domain experts are short, the sheer number of new objects is overwhelming. In turn this may lead to severe quality problems considering typical metadata quality measures like completeness, correctness and relevance. Thus, today libraries find themselves in the unfortunate position to either sacrifice metadata completeness when using traditional (mostly manual) indexing methods, or to use automatic techniques for metadata generation, which may have a negative impact on metadata correctness. Moreover, the question of metadata relevance today is strongly influenced by anticipating the possible usage of a resource, which in times of highly interdisciplinary research problems poses quite a difficult problem.

Basic techniques for automatic metadata generation can be roughly divided in text mining approaches, where a set of linguistic machine learning techniques are used mainly for natural language understanding, and automatic metadata extraction, where statistical machine learning techniques are used for specialized tasks like recognition and disambiguation of named entities, events, etc. Of course there often is a rather blurry line between individual techniques in this wide field of approaches.

The central point for most text mining approaches is to model, structure and annotate the meaning of textual sources within the context of a discipline. The aim is to create some degree of intelligent understanding that can subsequently be used for sophisticated tasks like document classification (e.g., distinguishing content from different disciplines or linking content to categories in taxonomies or ontologies), sentiment analysis (e.g., in user comments or scientific discussion forums), or establishing document similarity (e.g., in the sense of research using similar methodologies, working towards similar aims, etc.). Still, although today some tasks can already be solved within an adequate precision, deeper

investigations of automatic text classifications often show poor results for some cases that are bound to affect the overall mining performance and call for rigid (usually manual) quality control; see for instance (Barthel, Tönnies & Balke, 2013) for quality levels in a use case of mathematical document classification.

In contrast to the actual understanding of language the idea of automated metadata extraction is to find and annotate characteristic entities that are important for individual library items. Depending on the discipline there may be different kinds of entities, which however need to be treated as first class citizens in order to allow for subsequently offering sophisticated library services, see e.g. (Pinto & Balke, 2015) for a thorough discussion. For instance, in life sciences gene, proteins, or enzymes need to be extracted and annotated, mathematical items mostly need support for formulae, substances and reactions are important for chemistry documents, and drugs and diseases in the field of medicine. It is easy to see that handling all these different extractions needs a variety of methods as well as high expertise in machine learning plus deep domain knowledge. But recent research also shows that a high-quality extraction of relevant entities offers chances for additional descriptive annotations repurposed from semantically similar, yet often more general fields. Thus new access paths to information become available for subsequent use in value-adding services, see (Köhncke, Siehndel & Balke, 2015) for an example in chemical document collections.

## Using Social Sources for Content Enrichment

In times of the information deluge, the enrichment of content also needs to be performed in the sense of assessing the value of some library item as opposed to similar items. While in the last section we only focused on content-based value, in this section we will focus on the value expressed by previous users' experiences with the object. This assessment does not only focus on the filtering of erroneous (or even fraudulent) results, which are part of the self-correcting nature of science, although currently an increasing number of retractions can be observed. It is more about assessing a specific objects usefulness in the general sense, i.e. its perceived popularity given arbitrary intentions of use. Since interaction will of course often be based on a certain user intention and topical focus, social assessments today are often used as tie-breakers when too many relevant objects are retrieval.

Given the necessary trade-off between a researcher's time restrictions and increasing numbers of publications in his/her field, bibliometrics or scientometrics already play an important role today. Citation counts to assess the quality of individual publications, the h-factor to assess some author's prolificness or techniques like bibliographic coupling for deriving topical similarity between publications have been used for some decades. Despite many critical discussions and the showcasing of potential dangers for the scientific system (see for example the bluntly put, if somewhat simplifying 'weapons of mass citation' (Molinié & Bodenhausen, 2010) or 'impact factor wars' (Brumback, 2009)), such metrics are heavily used in practice. One example are metric-based Web retrieval services like Google Scholar or Microsoft Academic Search, another example are creating transparent metrics for candidate rankings in hiring procedures or tenure decisions at many universities and research institutions.

As opposed to the citation of previous work via new publications, an alternative source for social assessment of popularity are direct interaction measures. Such direct metrics include download numbers and page visits, DOI hashtags in tweets, 'likes' on platforms like Facebook, Digg or CiteULike, inclusion in Web-based reference managers lite Zotero and Mendeley, or user comments and discussions on dedicated Web forums. As stated in the manifesto (Priem, Traborelli, Groth & Neylon, 2010) of the altmetrics group, altmetrics in a way 'crowdsources' peer review by taking user interactions as votes of confidence for scientific content. Still, whether altmetrics can actually add a valuable facet for the design of helpful library services is an ongoing and open discussion (Barthel, et al., 2015).

## The Case for Contextualization and Personalization of Library Services

With today's development of more and more complex information retrieval and knowledge discovery scenarios the need for effective contextualization and personalization of library services has become a

demanding issue. Rather than using traditional library services the users' notion has changed to accomplishing tasks, i.e. not the service itself, but rather the task that it was designed for, is considered to be important. Thus, recently service designers have tried to improve the interactive behavior of services and gradually adapt it to the situation of interaction between humans. Technologies like navigation path tracking for user modeling and even the interpretation of direct interaction signals like mouse movements or physical expressions (eye movement, facial expressions) have become popular to assess user experience in real-time.

Personalization for services mainly consists of two parts. There is basic, usually implicit knowledge that can be gathered from the domain when offering a service (contextualization) and there is data that has to be user-provided for adequate content- and service-selection (personalization). As an example for the case of chemical digital libraries see the contextualization of document retrieval services in (Köhncke,Tönnies & Balke, 2012) as opposed to the personalization of document retrieval services in (Tönnies, Köhncke & Balke, 2011). While the contextualization focuses mostly on the enriched information of library objects and typical workflows in the domain, the necessary information for effective personalization is usually hard to come by. Here, different users' notions, conceptions, and conceptualizations form individual knowledge spaces that are hard to extract, but will strongly impact necessary tools for building services like e.g., deriving relevance-based rankings, measuring similarity between entities, or expressing the relatedness of concepts.

In any case, knowing the exact task that the user is about to perform has some advantages. Not only can the user provided data be used in a more effective way and the set of necessary data be minimized, but also the knowledge that is implicit in the task can be used at a finer granularity. In particular, sophisticated heuristics that may only apply to this special case can be applied and subsequently a better quality of service can be reached. Knowledge of the task at hand can be gained from knowing a user's intentions. However, the task that is about to be performed using a ready-made service may still somewhat differ from the user's actual intentions of what task to accomplish. In the worst case the service used may even be inappropriate to match the intention at all. Yet, when looking at practical user interactions with library services, also creative use in the sense of repurposing can often be observed.

## Conclusions

In this paper we have looked at some of the challenges that university libraries are currently faced with. In particular, the change from knowledgeable service providers in different disciplines to actual research institutions proactively combining knowledge from different domains and a thorough understanding of user experience and innovative business models, with state of the art digital techniques in novel and yet unforeseeable ways.

University libraries thus have to move into a new role with new and difficult tasks. Service design, evaluation, maintenance, and refinement will be constant processes that in future decide about a library's service provisioning qualities and to some degree also about its competitiveness in an increasingly diverse market for knowledge providers.

Finally, taking a closer look at the indexing process, which forms the backbone of library services, we argue that besides truly semantic annotations of larger structures (e.g., full texts, primary research data, or audiovisual content), also vital entities in each field have to be annotated and enriched on a fine-granular level. But as exemplified above, treating such entities as first class citizens needs a deep understanding und what is more an active, experience-driven evolution of digital techniques, often across disciplinary boundaries. Moreover, deriving information from different sources like content-based information, social information and information gained from interaction promises to alleviate some problems of the information deluge.

**References**


Barthel, S., Tönnies, S., and Balke, W.-T. (2013). Large-Scale Experiments for Mathematical Document Classification. *15th International Conference on Asia-Pacific Digital Libraries (ICADL)*, Bangalore, India.

Barthel, S., Tönnies, S., Köhncke, B., Siehndel, P., and W. - T. Balke (2015). What does Twitter Measure? Influence of Diverse User Groups in Altmetrics. *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Knoxville, TN, USA.

Brumback, R., (2009). Impact factor wars: Episode V—the empire strikes back. *Journal of Child Neurology*, Vol. 24(3).

Köhncke, B., Siehndel, P., and Balke, W.-T. (2015). Bridging the Gap - Using External Knowledge Bases for Context-Aware Document Retrieval. *15th International Conference on Asia-Pacific Digital Libraries (ICADL)*, Bangalore, India

Köhncke, B., Tönnies, S., and Balke, W.-T. (2012). Catching the Drift – Indexing Implicit Knowledge in Chemical Digital Libraries. *International Conference on Theory and Practice of Digital Libraries (TPDL)*, Paphos, Cyprus.

Kümmel, C., Reinhardt, A. (2011). Information Services of the Future: What is the Contribution of Special Subject Collections in German Libraries? *DFG-Infobrief: Research Funding - Facts and Figures*

Molinié, A. and Bodenhausen, G. (2010). Bibliometrics as weapons of mass citation. *Chimia (Aarau)*. Vol. 64(1-2).

Pinto, J. M. G. and Balke, W.-T. (2015). Demystifying the Semantics of Relevant Objects in Scholarly Collections: A Probabilistic Approach. *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Knoxville, TN, USA.

Priem, J., Taraborelli, D., Groth, P., and Neylon, C. (2010). Altmetrics: A manifesto. *http://altmetrics.org/manifesto (last accessed 15.08.2015).*

Tönnies, S., Köhncke, B., and Balke, W.-T. (2011). Taking Chemistry to the Task – Personalized Queries for Chemical Digital Libraries. *11th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Ottawa, Canada.

Ware, M. and Mabe, M. (2015). The STM Report. *International Association of Scientific, Technical and Medical Publishers (STM)*.