Wolfram Horstmann* und Jan Brase

Libraries and Data – Paradigm Shifts and Challenges

DOI 10.1515/bfp-2016-0034

Abstract: In a changing world, libraries have not only to keep in step with the developments but also to exert their influence as far as possible. Relevant developments congregate in the transformation of the library as a physical concept into a central learning and gathering place at the research campus, illustrating that the functional concept of a library is needed more than ever. As trusted advisers for the academia libraries can play a fundamental role in a time where more and more scholars are afraid of "information overload" and the definition of academic resources includes increasingly heterogeneous content types and new definitions of publishing. The new nature of tasks and content types ask for new services and new tools. The development and reliable maintenance of such services and tools are genuine tasks for libraries.

Keywords: Libraries; transformation; paradigm shifts; research data; information hub

Bibliotheken und Daten – Paradigmenwechsel und Herausforderungen

Zusammenfassung: In einer sich verändernden Welt müssen Bibliotheken nicht einfach nur mit den Veränderungen Schritt halten, sondern sie – soweit möglich – aktiv mitgestalten. Die Bibliothek wandelt sich von einem physischen Ort des Wissens hin zu einem zentralen Lern- und Kommunikationsort am Forschungscampus, aber bei allem bleibt das funktionale Konzept einer Bibliothek bedeutsam. Als vertrauenswürdige Berater der Forschung kommt den wissenschaftlichen Bibliotheken eine fundamentale Rolle zu, insbesondere vor dem Hintergrund des „information overload" und zunehmend heterogenen Datentypen und einer sich verändernden Publikationskultur. Der neue Charakter dieser Aufgaben und Datentypen verlangt nach neuen Diensten und Werkzeugen, deren Bereitstellung und Weiterentwicklung eine natürliche Aufgabe für Bibliotheken sind.

***Corresponding author: Wolfram Horstmann,** horstmann@sub.uni-goettingen.de
Jan Brase, brase@sub.uni-goettingen.de

Schlüsselwörter: Bibliotheken; Transformation; Paradigmenwechsel; Forschungsdaten

1 The forth paradigm and the role of libraries

The concept of a "Library" has changed dramatically through the last years. This can best be described in parallel to what Jim Gray introduced as the four paradigms of science.[1] Gray points out that since a thousand years research is intellectual and observational, describing natural phenomena. The last few hundred years then saw a theoretical branch evolving, followed by models and generalizations were used to understand what was behind these natural phenomena, thus making the shift from the first paradigm to the second, with the scholars no longer being a passive observer, but actively trying to find out, why things are like they are.

Accompanying this development, libraries for centuries were a major, if not the main research infrastructure for academic institutions. They started off by holding the manuscripts and prints of researchers working at the institution, in times when reproduction of scholarly work was the exception and scholars had to travel around the world to gain insights into the works of other scholars. When the reproduction of scholarly works became easier, libraries were able to collect a large segment of the world's knowledge and make it accessible to researchers and students. Libraries' estates were usually established at the heart of the campus to perform their organizational function for the circulation of knowledge and serve as a sanctuary for study. This traditional function of a library has been dominated by "books" – hence the term Library, from Latin 'liber'. However, by no means books were the only things in traditional libraries: they held maps, manuscripts, paintings, and archives of all kinds, even ephemera like shoes.

The digital revolution in the last few decades allowed a computational branch to grow with the opportunity to use the developed theories to simulate complex phenom-

1 Hey (2009).

ena. This was of course the shift from the second paradigm to the third, allowing the scholars to simulate reality in order to test in detail their theory against their empirical observation. This digital age amplified the role of libraries as multi-data institutions, specifically in academic libraries. Electronic catalogues were invented in the 50's and introduced as enterprise systems as early as in the 70's. In the meantime, e-journals and e-books are often more frequently used in libraries than physical books and licensed databases are longstanding regular services offered to academics. Libraries are operated as a highly digital enterprise. They do not only run extremely complex and internationally networked metadata systems but also a highly demanding circulation and licensing business for physical and electronic resources – often for 10 thousands of individually registered users. And this is at only one institution. All library data put together would probably form the most complete record of academic work worldwide. It is for this standing in their institution that libraries often also run other business for the university such as identity management or multi-purpose university cards, and they increasingly take over research information management, i. e. authoritative records of persons, publications, projects, and organisational units.

Today, with the next paradigm shift, we encounter what is labelled as enhanced science or eScience: Data intensive research that unifies theory, experiment, and simulation. This is what Jim Gray defined as the fourth paradigm.

Now, why is this important for libraries and what are the consequences of this for them? Firstly, libraries have a strong mandate of offering access to information and knowledge. Secondly libraries have a long tradition and experience in providing and curating information, as they are doing so for thousands of years. This makes libraries trustworthy organizations that also have a role in society to be persistent. Especially in the digital age, where more on more information is only available in electronic formats this has become more and more important. While there is always a great risk that current projects and initiatives that create information will no longer be around after a decade, or to be more precise after the funding stops, the chance that the libraries will still be around is considered a given fact.

Following the paradigm shifts, information nowadays is more than articles or books or any kind of mostly textual information. For libraries to keep a record of research, libraries need to keep a record of other things than books. And libraries are already a digital data powerhouse. If all this was true, why do libraries do not keep a systematic record of research *data*?

Obviously part of the answer lies in the complexity and high context-specificity of research data. The context of data is characterized by the subject of research – the discipline – the phase of the individual research process – from data generation to data publishing and re-use – the funding, privacy and copyright regime and many other 'soft' characteristics, among them the temper and ideology of the researchers involved. In a research cycle, the researcher creates various versions of data sets, which often are recorded in the same database or repository. The data set is therefore a composite object. The identifying descriptors of that object must include enough specificity about its constituent parts so that a scholar can refer to one and only one, unambiguous, clearly defined data set. This requires versioning of records and identification of entities that have contributed to the data or changed them, differentiating such details as the author for data creation from the author of data interpretation. However, research data management is not only a problem of assigning identifiers or metadata. For the purposes of aggregation, computation, verification, reproducibility, and replicability, the data set must be defined so that it can be referenced in a way that yields a concrete search result.[2] This complexity is one of the reasons, why libraries do not yet keep a systematic of research data.

2 The Historical Separation of Libraries and Data through Labs

If we go back 350 years into the 17th century, we will find an intensive discussion on the fundamental question of how to do research was going on. There was a group of scholars claiming that "the word" is not enough to make statements about how the world works. What you would need is evidence. It led to the foundation of the Royal Society in 1660. The motto in the coats of arms says "Nullius in verba": take no one's word for it. It is considered to be the birth of experimental research, gathering evidence for scientific claims, gathering research data, bringing us into the second paradigm. It is interesting to note that the Royal Society founded a journal, the "Philosophical Transactions", to share the results of this research. They also built two infrastructures: a library and a repository of specimen where evidence could be collected. However, somehow the good intention to keep a record of the evidence that is underlying the publication was lost. Labora-

2 Wynholds (2011) S. 215.

tories were built around the world and the evidence was kept in the laboratories while the record of research was 'only' kept in the text publications, which, in turn is kept in libraries – until today. Thus, the separation of libraries and labs had a strong impact. In our modern views on science policy this would not be labelled as "good scientific practice" of "proper conduct of research".

But something else, often unnoticed, happened through the new digital possibilities. The *term* library was transferred to a completely different context, most interestingly expressed by the concept of a *software* library. Here, the term library is today used completely dissociated from the texts and books, the 'soul' of the traditional library. It refers only to the main *function* of a library, namely an organized way of managing knowledge resources. At the same time libraries transcended their physical form in putting their printed holdings into a digital form through digitization.

This change led to the final establishment of two distinct concepts of a library:

- referring to a *physical* structure with walls and shelves for predominantly *texts and books*, and
- referring to a *functional*, maybe virtual, system for the *organisation of knowledge resources*.

This distinction can also be found in the use of the term "*Digital Library*": referring to the collection of digitized books in libraries, often related to unique or special collections and also referring to a field in computer science that has little resemblance to libraries (and maybe little resemblance to computer science, too).

3 Data Libraries

"Data Libraries" is a rather new term and has not yet been specified properly. Following the distinction introduced above, there are of course also two classes of "data libraries":

On the one hand data library is referring to library organisations, specifically academic libraries. In this sense, the data library is the above mentioned 'data powerhouse', i. e. the library as a business data organisation and a content-intensive organisation. The second class of a data library, i. e. the abstract, refers to collections of data that apply principles for the organisation of knowledge.

The term "Research Data Library" has two general meanings and diverse. It is too early to decide at this early stage of the emergence of data libraries about wrong or right, better or worse. But it is important to consider the spectrum:

Data Libraries emerging from the traditional physical world might have high competences in running business for knowledge organisation for 100s of years but might not have the skills to understand the complex nature of research data. Conversely, Data Libraries emerging from the virtual world might be agile and adaptive to researchers' needs but might have no idea how to run a sustainable organisation for more than 10 or 20 years. The growing need for the establishment of some form of data libraries however is compelling.

Knowledge, as published through academic literature, often is the last step in a process originating from research data. These data are analysed, synthesised, interpreted, and the outcome of this process is generally published in its result as a scholarly article.

Only a very small proportion of the original data are published in conventional academic journals. Existing policies on data archiving notwithstanding, in today's practice data are primarily stored in private files, not in secure institutional repositories, and effectively are lost.[3]

This lack of access to research data is an obstacle to international research. It causes unnecessary duplication of research efforts, and the verification of results becomes difficult, if not impossible.[4] Large amounts of research funds are spent every year to re-create already existing data.[5] Progress in sharing of research data has been made at a fast pace. Infrastructures such as grid exist for storage. Methodologies have been established by data curation specialists to build high quality collections of datasets. These include standards for metadata (provenance, copyright, author of a dataset), registration, cataloguing, archiving, and preservation. A large number of disciplines benefit from these methodologies and high quality datasets.

The network DARIAH as a Pan-European infrastructure for Arts and Humanities has always supported local data stores as data libraries for the trustworthy management of research data from the Humanities; this includes large national data archives as well as smaller specialised collection. At University of Goettingen the project "Humanites Data Center" (HDC) established a roadmap for the design of data centers that form a data library for the Humanities. The joint project was funded by the states of Lower Saxony and Berlin and was a joint approach of the library, the academies of science from Goettingen and

3 Lawrence and Krovetz (2001).
4 Dittert, Diepenbroek and Grobe (2001).
5 Arzberger, Schroeder, Beaulieu, Bowker and Casey (2004).

Berlin-Brandenburg and scientific computing centers in Goettingen and Berlin.

More generally, the project re3data.org offers a global registry of research data repositories and offers a broad overview on existing research data repositories from different academic disciplines. The registry went live in autumn 2012 and is funded by the German Research Foundation (DFG).[6]

4 Libraries as information hubs

The development of the internet in the last decades and the principle of linking content independently from its physical location dramatically changed the definition of a library catalogue. Traditionally a library catalogue has been seen as a window to the library's holdings, a structured summary of what can be brought easily to the shelf. Due to the growth of the internet in the last decades, this has slowly changed and more and more catalogues offer direct access to pdf-versions of documents, but the principle has been the same throughout the centuries.

Now in the fourth paradigm it becomes more and more impossible for a library to actively store all these kinds of information that are important for its users. Nevertheless, the great chance with the growth of the internet is that the library does not have to store this information, when it is available somewhere else in the internet. The libraries job in the future might be to know where the information is, if the content provider is trustworthy and to have a distinguished description of the content in its catalogue to offer the service of answering queries from user. In a nutshell, the library of the future should be able to answer the query of a user with the reference service that says, in expressed form: "We do not have what you are looking for, but we now where it is, and we can offer you a link to it". This implies many aspects: The library has to able to understand what the user is looking for. It has to able to have enough distinguished information about content in its catalogue to know what ideal results would be for the query. The library has furthermore to know where this content is stored and has to provide a persistent link to it.

Library catalogues are classical sources for information.[7] As explained earlier, when querying for a certain topic, users might not be interested in only receiving all relevant publications as a result, but also additional datasets collected by the corresponding scholar.

The assignment of persistent identifiers allows this research data to become directly accessible through library catalogues. When the persistent identifier of the dataset is resolved, the user does not directly download megabytes of data but is linked to a preview page where the data centre provides metadata and download links to different parts of the data. This workflow is similar to the use of Digital Object Identifier (DOI) names in scholarly journals, where the resolution of a DOI name of an article directs you to a publisher's page, including the metadata of the article.

The Digital Object Identifier DOI was introduced in 1998 with the funding of the International DOI Foundation (IDF). It is a registered trademark and DOI names can only be assigned by official DOI registration agencies that are a member of IDF. There currently are a total of 10 Registration agencies worldwide.

The DOI system is technically based on the non-commercial handle system of the Corporation for National Research Initiatives (CNRI). Since 2012, the DOI system is an official ISO standard (ISO 26324). Registration agencies are responsible for assigning identifiers. They each have their own commercial or non-commercial business model for supporting the associated costs. The DOI system itself is maintained and advanced by the IDF, controlled by its registration agency members.

DOI names from any registration agency can be by default resolved worldwide in every handle server; DOI therefore are self-sufficient and their resolution does not depend on a single resolution server. A standard metadata kernel is defined for every DOI name. Assigning DOI names involves the payment of a license fee by the Registration agency but their resolution is free.

DOI has emerged as the most widely used standard for digital resources in the publication world. It is currently used by all major scientific publishers and societies (Elsevier, IEEE, ACM, Springer, Wolters Kluwer International Health & Science, New England Journal of Medicine, etc.).

The international association DataCite was founded in 2009 to actively assign DOI names to research data sets, to allow research data to be handled and cited as independent published objects. DataCite is operated through its current 30 members, most of which are libraries from all over the world. With the assignment of over 7 million DOI names to research objects until now, DataCite has created the technical backbone to link to research data and other information objects directly from existing library catalogues.[8]

6 Pampel, Bertelmann, Scholze, Kindling and Vierkant (2004).

7 Inger and Gardner (2008).

8 Brase and Sens (2015).

In Germany the registration of DOI names for data sets is offered as a service for academia through five German libraries and based on the division of disciplines. The German National Library of Science and Technology (TIB) are offering the service for technology and fundamental sciences, the German National Library of Medicine (ZB Med), the German National Library of Economics (ZBW) and the Leibniz Institute for the Social Sciences (GESIS) for their corresponding disciplines. The Goettingen State and University library is offering the service for the Humanities in Germany.

The Goettingen DOI registration service is an integral part of the DARIAH infrastructure and available for every researcher or institution in Germany that wants to publish research data from the humanities.

5 The road ahead

As the world is changing, libraries have to change with it and actively change it, too. Whereas the physical concept of a library might evolve to a central learning and gathering place at campus, the functional concept of the library is needed more than ever. Libraries have a millennia spanning tradition of being keepers and finders for scholarly information. As trusted advisers for the academia libraries can play a fundamental role in a time where more and more scholars are afraid of "information overload" and the definition of academic resources includes increasingly heterogeneous content types and new definitions of publishing.

This is, thus, a challenging task but history has shown that libraries always have been able to adapt to these paradigm shifts. The new nature of tasks and content types ask for new services and new tools. The development and reliable maintenance of such services and tools are genuine tasks for libraries.

References

Arzberger, P.; Schroeder, P.; Beaulieu, A.; Bowker, G.; Casey, K. (2004): Promoting Access to Public Research Data for Scientific, Economic, and Social Development. In: *Data Science Journal,* (3), 135–52. Verfügbar unter https://pdfs.semanticscholar.org/5866/1dccf9c996e8ab0bb0afd1a5455e14ed1a99.pdf.

Brase, Jan; Sens, Irina (2015): The Tenth Anniversary of Assigning DOI Names to Scientific Data and a Five Year History of DataCite. In: *D-Lib Magazine,* 25 (1, 2). Verfügbar unter http://www.dlib.org/dlib/january15/brase/01brase.html.

Dittert, Nicolas; Diepenbroek, Michael; Grobe, Hannes (2001): Scientific data must be made available to all. In: *Nature,* 414 (6862), 393. doi: 10.1038/35106716.

Hey, Anthony J. G. (2009): The fourth paradigm: Data-intensive scientific discovery. Redmond, Wash: Microsoft Research.

Inger, Simon; Gardner, Tracy (2008): How readers navigate to scholarly content: Comparing the changing user behaviour between 2005 and 2008 and its impact on publisher web site design and function. Abingdon: Simon Inger Consulting.

Lawrence, Steve; Krovetz, Robert (2001): Persistence of web references in scientific research. In: *IEEE Computer,* 34 (2), 26–31. Verfügbar unter http://www.fravia.com/library/persistence-computer01.pdf.

Pampel, H.; Bertelmann, R.; Scholze, F.; Kindling, M.; Vierkant, P. (2004): Stand und Perspektive des globalen Verzeichnisses von Forschungsdaten-Repositorien re3data.org. In: In Müller, P; Neumair, B.; Reiser, H.; Rodosek, G. D. (Eds.): *8. DFN-Forum Kommunikationstechnologien. Beiträge der Fachtagung. Lecture Notes in Informatics (LNI) – Proceedings*. Series of the Gesellschaft für Informatik (GI). Volume P-243. Bonn: Gesellschaft für Informatik, 13–22.

Wolfram Horstmann
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Der Direktor
Platz der Göttinger Sieben 1
D-37073 Göttingen
horstmann@sub.uni-goettingen.de

Jan Brase
Niedersächsische Staats- und Universitätsbibliothek Göttingen
Abteilung Forschung und Entwicklung
Platz der Göttinger Sieben 1
D-37073 Göttingen
brase@sub.uni-goettingen.de