

# Schluss mit den ewigen Fragen? Langzeitverfügbarkeit als exemplarische Aufgabe der Informationsinfrastruktur

Frank Scholze (frank.scholze@kit.edu), Karlsruher Institut für Technologie (KIT), KIT-Bibliothek

Wissenschaftliche Bibliotheken haben viele Baustellen – angefangen von ihrer grundlegenden Stellung und Bedeutung in einer digitalen oder zumindest sich rasch digitalisierenden Wissenschaft und Gesellschaft über Fragen der Organisation, des Personaleinsatzes und der Finanzierung bis hin zur Entwicklung neuer Dienstleistungen in den Bereichen Forschungsdaten, elektronisches Publizieren oder Soziale Netzwerke.<sup>1</sup> Erstaunlich ruhig ist es um das einst heiß diskutierte Thema der digitalen Langzeitarchivierung bzw. Langzeitverfügbarkeit geworden.<sup>2</sup> Ist das Problem also gelöst? Es existiert ein allgemein akzeptiertes Konzeptmodell (OAIS), es existieren Softwarelösungen im produktiven Einsatz (u.a. an der BSB München, der SLUB Dresden und auch an der ETH Zürich) und es gibt in Deutschland das Kompetenznetzwerk nestor.<sup>3</sup>

## Organisation

Irritierend mag einem dabei höchstens vorkommen, dass die Deutsche Forschungsgemeinschaft zwei Ausschreibungen zu dieser Thematik unternahm, die ergebnislos blieben. Auf der Basis des DFG-Positionspapiers „Weiterentwicklung der Bibliotheksverbände als Teil einer überregionalen Informationsinfrastruktur“ hatte die DFG am 12. Januar 2012 die Ausschreibung „Neuausrichtung überregionaler Informationsservices“ veröffentlicht.<sup>4</sup> Sie zielte darauf ab, einen umfassenden Umstrukturierungsprozess anzustoßen, in dessen Verlauf die derzeit vorrangig regional orientierten Informationsangebote zu nationalen, aber auch international vernetzten, funktional definierten,

---

<sup>1</sup> Exemplarisch Achim Bonte (2014): Wissenschaftliche Bibliotheken der nächsten Generation. Sind die Institutionen und ihre Mitarbeiter für die Zukunft gerüstet? In: ZfBB Vol. 61, No. 4/5, S. 239-242, <http://dx.doi.org/10.3196/18642950146145114>

<sup>2</sup> Jeff Rothenberg (1995): Ensuring the Longevity of Digital Documents In: Scientific American Vol. 272, No. 1, S. 42-47

<sup>3</sup> <http://www.langzeitarchivierung.de>. Die folgenden Betrachtungen beziehen sich im Wesentlichen auf Deutschland.

<sup>4</sup>

[http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung\\_ueberregionale\\_informationsservices\\_121015.pdf](http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_ueberregionale_informationsservices_121015.pdf)

nachhaltigen Services weiter entwickelt werden. Die Ausschreibung sprach explizit das Themenfeld „Langzeitverfügbarkeit“ an. Dieses erbrachte als einziges der in der Ausschreibung benannten Themenfelder kein förderfähiges Projekt; daher wurde diese Aufgabenstellung am 15. November 2013 erneut ausgeschrieben – wiederum ohne Ergebnis.<sup>5</sup> Eignet sich das Thema möglicherweise nicht für strukturelle Reformen, die ja das erklärte Ziel der ersten Ausschreibung war? Betriebswirtschaftlich gibt es kaum ein Thema in Bibliotheken oder Archiven, das eine schlechtere Kosten-Nutzen-Relation aufweisen würde als digitale Langzeitverfügbarkeit. Volkswirtschaftlich, kultur- und gesellschaftspolitisch ist der Nutzen beinahe unermesslich, aber doch indirekt, da nicht jederzeit und für jeden sofort fühlbar, sondern erst ex negativo definiert. Gerade darum bietet sich dieses Handlungsfeld für Strukturreformen an wie kein zweites. Nur eine kleine Gruppe leistungsfähiger Institutionen kann dieses Spannungsfeld aus betriebs- und volkswirtschaftlicher Betrachtungsweise erfolgreich bestehen und bestreiten. Auch eine Raumstation wird nicht jeder bauen können und wollen. Daher bleibt die Frage, ob die Selbststeuerungsfähigkeit im Bereich der Informationsinfrastruktur in Deutschland nur zu Kompetenznetzwerken reicht, d.h. dem kleinsten gemeinsamen Nenner. Oder sprechen wir lieber von einem Ökosystem, bei dem zwar alle Teile in einer Beziehung zueinander stehen, aber auch die Mechanismen der Evolution (Anpassung und Selektion) gelten? Nicht umsonst betont Norbert Lossau im Vorwort zum Sammelband *Evolution der Informationsinfrastruktur*<sup>6</sup> die Notwendigkeit der engen Partnerschaft bei gleichzeitigem Herausbilden vorteilhafter Eigenschaften.

Die Aufgaben der Informationsinfrastruktur für die Wissenschaft werden zunehmend komplexer, während die Umgebungstechnologien und Konzepte eine rasche Entwicklung durchlaufen. Dies kann nur in größeren Partnerschaften bzw. Zusammenschlüssen bewerkstelligt werden. Hier ist die Evolutionsmetapher auf die Herausbildung komplexer Organismen zu beziehen, die enger und intensiver interagieren und damit erfolgreicher agieren. Damit ist die Grenze des Vergleiches jedoch ebenfalls aufgezeigt. Menschen und Institutionen sind an sich bereits komplexe Organismen und es sind hier zwei grundlegende Fragen zu beantworten. Wie eng binden sich Akteure der Informationsinfrastruktur aneinander und wie groß ist die Fertigungstiefe, d.h. der Grad der Externalisierung – kurz gesagt, wer ist wie Teil des Gesamtsystems und wo verlaufen die Grenzen „nach außen“?<sup>7</sup> Trotz vielfältiger fachlicher und politischer Aktivitäten in den vergangenen Jahren, die 2014 mit der Ein-

---

<sup>5</sup>

[http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung\\_ueberregionale\\_informationssservices\\_131115.pdf](http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_ueberregionale_informationssservices_131115.pdf)

<sup>6</sup> Evolution der Informationsinfrastruktur: Kooperation zwischen Bibliothek und Wissenschaft (2013). Heike Neuroth, Norbert Lossau, Andrea Rapp (Hrsg.) Glückstadt: vwh-Verlag. <http://dx.doi.org/10.3249/webdoc-39006>.

<sup>7</sup> Grundlegend zur Frage der Organisationsform von Akteuren in Netzwerken und zur Fertigungstiefe Jörg Sydow (1992): Strategische Netzwerke : Evolution und Organisation, Wiesbaden : Gabler, S. 103ff.

setzung eines Rates für Informationsinfrastrukturen durch die Gemeinsame Wissenschaftskonferenz (GWK)<sup>8</sup> eine sichtbare strukturelle Ausprägung erhalten, sind wir immer noch erst am Beginn des Weges, die oben gestellten Fragen zu beantworten. Es wäre verfehlt, vom Rat für Informationsinfrastrukturen zu erwarten, nun jegliche Strukturreform oder die Ordnung des Ökosystems leisten zu können. Er sollte sich jedoch durchaus soweit von sachfremden politischen Erwägungen frei machen, dass er als Evolutionskatalysator wirkt und einen Kulturwandel zu einem stärker interorganisationalen Netzwerk einleitet, das gleichsam wie „ein Unternehmen“ agiert. In Bezug auf das Thema Langzeitverfügbarkeit bedeutet dies, engere Partnerschaften oder auch Integrationen relevanter Akteure zu entwickeln, offener mit Fehlern und Ungelöstem umzugehen und Konzepte und Technologien nochmals zu überdenken.

## Konzept

Martin Klein und seine Kollegen vom Digital Library Research and Prototyping Team am Los Alamos National Laboratory<sup>9</sup> untersuchten vor wenigen Monaten einen Korpus von rund 3,5 Millionen STM-Zeitschriftenaufsätzen aus den Jahren 1997 bis 2012 auf die Verfügbarkeit der referenzierten Webquellen. Dabei prüften sie nicht nur die Originalquelle, sondern auch eine Reihe von Langzeitverfügbarkeitssystemen wie dem British Library Web Archive u.a. Als Resultat zeigte sich, dass ein Fünftel der Aufsätze nicht mehr auflösbare Referenzen enthielten. Dies ist unter dem Gesichtspunkt der Nachvollziehbarkeit und Reproduzierbarkeit wissenschaftlichen Arbeitens ein katastrophaler Befund. Natürlich könnte man jetzt argumentieren, dass, wenn womöglich alle Archive der Welt mit eingebunden gewesen wären, dieser Anteil noch geringer wäre. Diese Aussage ist jedoch Teil des konzeptionellen Problems. Elektronische Langzeitverfügbarkeitssysteme und ihre Inhalte müssen, wenn sie schon - aus Sicherheitsgründen o.ä. - nicht direkt zugreifbar sind (dark archive), so doch zumindest in der digitalen Infrastruktur bekannt sein. Bekannt bedeutet dabei nicht intellektuell, sondern maschinell. Das hiberlink-Projekt<sup>10</sup> gibt eine mögliche Antwort darauf, indem es versucht, Langzeitverfügbarkeitssysteme in wissenschaftliche Prozesse zu integrieren, z.B. in Literaturverwaltungsprogramme, die beim Schreiben eines Artikels eingesetzt werden. Wie stark diskutieren wir in Deutschland das Zusammenwirken aller

---

<sup>8</sup> <http://www.gwk-bonn.de/themen/uebergreifende-wissenschafts-und-forschungspolitische-themen/informationsinfrastruktur/>

<sup>9</sup> Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, et al. (2014): Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. In: PLoS ONE 9(12): e115253. DOI: 10.1371/journal.pone.0115253

<sup>10</sup> <http://hiberlink.org/developments.html>

LZA-Systeme im Hinblick auf ihre Einbindung in digitale wissenschaftliche Workflows? Oder genügt uns die Deutsche Digitale Bibliothek?

## Technik

Gordon Moore, Mitbegründer von Intel, beschrieb seine Beobachtung, dass sich Speicherkapazität annähernd exponentiell entwickeln werde, bereits 1965.<sup>11</sup> Die Bezeichnung „Moore'sches Gesetz“ wurde um 1970 von Carver Mead, einem US-amerikanischen Pionier der modernen Mikroelektronik geprägt. Heute geht man davon aus, dass sich alle 18 Monate die Datenmenge verdoppelt, die sich auf einem gängigen Festplattenlaufwerk speichern lässt. Speichertechnologien werden sich in den nächsten Jahren grundlegend wandeln. Schon aus Performance-Gründen sind neue Wege gefragt. Denn die Steigerung der speicherbaren Datenmenge genügt nicht - Daten müssen auch in angemessener Zeit verarbeitet werden. Aus Sicht der traditionellen Wissenschaftskommunikation ist dies jedoch kein grundsätzliches technisches Problem. Auch bei einer Zuwachsrate von rund 3% pro Jahr und einer großzügigen Auslegung von Umfang und Artikelgröße benötigt der Gesamtkorpus aller wissenschaftlichen Aufsätze Speicherplatz im niedrigen dreistelligen Terabyte-Bereich.<sup>12</sup> Moderne Rechenzentren sind bereits in der Exabyte-Speicherung aktiv oder bereiten diese vor.<sup>13</sup>

Und es geht noch weiter. Wissenschaftler der ETH haben kürzlich erste Ergebnisse einer Speicherung digitaler Dokumente mittels DNA veröffentlicht.<sup>14</sup> Für diese Art der Speicherung sind, je nach Lagerungsbedingungen, experimentell Standzeiten zwischen 2000 und einer Million Jahre ermittelt worden. Im Vergleich dazu lassen sich Daten auf Mikrofilm nur für schätzungsweise 500 Jahre bewahren.<sup>15</sup> Um diese Dauerhaftigkeit zu erreichen, verkapselten die Forschenden die informationstragenden DNA-Stücke in Siliziumdioxid (Glas), zum anderen verwendeten sie einen Algorithmus, um Fehler in den ausgelesenen Daten zu korrigieren. Da der Einschluss in Siliziumdioxid ungefähr demjenigen in fossilen Knochen entspricht, konnten die Forschenden auf diese prähistorischen Daten über die Langzeitstabilität von verkapselter DNA zurückgreifen. Theoretisch könne ein einziges Gramm DNA 455 Exabyte, also

---

<sup>11</sup> Gordon E. Moore (1965): Cramming more components onto integrated circuits. In: *Electronics*, Vol. 38, No. 8, April 19, 1965

<sup>12</sup> Daten der World Bank. <http://data.worldbank.org/indicator/IP.JRN.ARTC.SC?page=1>

<sup>13</sup> <http://www.zdnet.com/article/worlds-1st-exabyte-storage-system/> Vgl. auch [bwDataArchiv http://www.scc.kit.edu/forschung/9645.php](http://www.scc.kit.edu/forschung/9645.php)

<sup>14</sup> Robert Grass, Reinhard Heckel, Michaela Puddu, Daniela Paunescu and Wendelin Stark (2015): Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. In: *Angew. Chem. Int. Ed.*, 54: 2552–2555. DOI: 10.1002/anie.201411378 <http://onlinelibrary.wiley.com/doi/10.1002/anie.201411378/pdf>

<sup>15</sup> Frauke Focke (2009): Datenspeicherung - Für die Ewigkeit. In *Spektrum der Wissenschaft* März 2009, <http://www.spektrum.de/news/fuer-die-ewigkeit/985077>

455 Milliarden Gigabyte an Informationen speichern, behaupten die Wissenschaftler.<sup>16</sup>

Wo also liegt das Problem? Darin, dass technische Invention noch keine Innovation im Sinne einer umfassend eingesetzten und erfolgreichen Anwendung darstellt.<sup>17</sup> Invention macht lediglich Potentiale sichtbar, zeigt mögliche Wege auf. Dies gilt für DNA-Speicher ebenso wie für das hiberlink-Projekt. Erst die Verbindung von Technik, Konzept und Organisation kann die Frage nach der Langzeitverfügbarkeit digitaler Informationen (ebenso wie andere komplexe Aufgaben der wissenschaftlichen Informationsinfrastruktur) lösen. Es liegt also einmal mehr an uns, den Akteuren in Wissenschaft, Informationsinfrastruktur und Politik, über den kleinsten gemeinsamen Nenner hinauszugehen und organisatorisch mehr zu schaffen als ein Kompetenznetzwerk oder technisch mehr als Insellösungen.

---

<sup>16</sup> Moritz Geier (2015) Ein Stück DNA kann digitale Daten für mehr als zwei Millionen Jahre speichern. In: Wired Februar 2015. <https://www.wired.de/collection/latest/ein-bisschen-dna-speichert-daten-besser-als-jede-festplatte>. Im Wesentlichen hängt die Speicherdichte von der Redundanz der gespeicherten Information ab, konservative Ansätze gehen von 700-750 Terrabyte pro Gramm aus. Vgl. George M. Church, Yuan Gao and Sriram Kosuri (2012) Next-Generation Digital Information Storage in DNA. In: Science 28 September 2012: 337 (6102), 1628. DOI:10.1126/science.1226355

<sup>17</sup> Joseph A. Schumpeter (1939): Business Cycles. A Theoretical, Historical, and Statistical Analysis of the Capitalist Process. New York: McGraw-Hill.