

Überlegungen zu einer nachhaltigen Forschungsdateninfrastruktur

Bernd Kulawik

In diesem Essay wird eine radikale Lösung des Problems wirklich langfristiger Datensicherung und -verfügbarkeit vorgeschlagen. Sie beruht auf der zuvor erläuterten Einschätzung, dass die Vielfalt und Komplexität heutiger IT-Systeme weder mit vertretbarem Aufwand für die Zukunft konserviert werden kann, noch in der Zukunft rekonstruierbar sein wird. Ein Ausweg könnte die Schaffung einer (internationalen) Institution mit "Ewigkeitsgarantie" sein, welche die Entwicklung und Betreuung einer freien Hard- und Softwareplattform übernimmt, mit der insbesondere Forschungsprojekte und -einrichtungen, aber auch Firmen oder Privatpersonen arbeiten können, die ihre Daten abschließend der Institution zur dauerhaften Speicherung übergeben wollen. Bei der Entwicklung dieser Plattform könnten vorausweisende Lösungen aus der bisherigen IT-Entwicklung übernommen; bekannte, aber aus Kompatibilitätsgründen nicht behobene Fehler und Irrtümer jedoch vermieden werden. Außerdem könnte das System zugleich als eine wirklich offene Plattform für Repositorien zur Veröffentlichung von Daten und Ergebnissen im Open Access bilden und undurchsichtige Verfahren beispielsweise im Peer Review und der Messung von Forschungsrelevanz ablösen. Da Bibliotheken, Archive und Museen diejenigen Institutionen sind, die über umfangreichste historisch gewachsene Erfahrungen in der Speicherung und Erschließung von Daten im weitesten Sinne haben, sollten sie bei Konzeption und Betrieb dieser Institution eine führende Rolle spielen.

Einleitung: Zustandsbeschreibung und abzuleitende Forderungen

Aufgrund sinkender Kosten für die Papierherstellung kam es in der ersten Hälfte des 19. Jahrhunderts zu einem so explosionsartigen Anstieg des Buchdrucks, dass keine einzelne Bibliothek mehr den Ankauf aller relevanten Literatur bewältigen konnte. Daraufhin entstanden Staats- und Nationalbibliotheken, deren Aufgabe die Dokumentation der gesamten in einer Sprache oder einem Land gedruckten Veröffentlichungen war und ist. Dazu wurden sie mit einer „Ewigkeitsgarantie“ versehen, das heißt einer langfristig gesicherten Unterhaltung aus Steuermitteln. Zugleich wurden zur Erschließung des gespeicherten Wissens normierte Katalogisierungssysteme entwickelt, die standardisiert als Vorbilder oder Vorgaben durch andere Bibliotheken genutzt werden konnten.

Hinsichtlich des heute ebenfalls exponentiell wachsenden digital gespeicherten Daten befinden wir uns meines Erachtens heute in einer vergleichbaren Situation. Nur ist aus der zweidimen-

sionalen Gutenberg-Galaxis des bedruckten Papiers ein drei-, eigentlich – wegen der zeitlichen Komponente – sogar vierdimensionales digitales Universum mit Milliarden von Galaxien geworden. Selbst wenn für die Zukunft wie bisher nur ein Bruchteil dieser Informationen aufbewahrt werden kann und soll, so ist allein deren Umfang immer noch so riesig, dass er selbst nationale Institutionen überfordern dürfte.

Hinzu kommt aber noch ein grundlegendes Problem, vor dem die Bibliotheken bisher nicht standen: Die tatsächliche Verfügbarkeitsdauer dieser digitalen Daten sowie der zu ihrer Benutzung notwendigen Software und selbst der Hardware tendiert aus der Sicht des Historikers gegen Null: Die von Forschungsförderorganisationen für die nun zunehmend – von den Wissenschaftlern! – eingeforderten „Forschungsdatenmanagementpläne“ zielen auf Zeiträume von 10-15, eventuell auf 20 und im Idealfall auf 50 Jahre ab... Aus dem Blickwinkel der Bibliothek, die jahrhundertealte Bücher und jahrtausendealte Papyri oder Tontafeln aufbewahrt, ist dies geradezu lächerlich...

Inzwischen gibt es erste Ansätze zur Entwicklung beispielsweise *Nationale Forschungsdateninfrastrukturen*. Dabei sollen durch Konsortien betroffener Institutionen Vorschläge für (eine) solche Infrastruktur(en) ausgearbeitet und in einem Wettbewerb ausgewählt werden, an dessen Ende dann ein System stehen soll, das geeignet sein müsste,

- die vorhandenen und zukünftigen Daten aus Forschungsprojekten und
- die zumeist projektspezifisch angepasste Software (zum Beispiel Datenbankschemata und Benutzerinterfaces) aufzunehmen sowie
- gegebenenfalls die dazugehörige Hardware zu simulieren.

Diese Bemühungen kommen eigentlich circa 25 Jahre zu spät, denn nicht nur das seit Anfang der 1970er Jahre unter der Bezeichnung Arpanet bestehende heutige Internet wurde vor allem durch die Wissenschaft genutzt. Auch das 1991 von Tim Berners-Lee am CERN vorgestellte World Wide Web (WWW) und sein grundlegendes HyperText Transport Protocol (HTTP) waren *in erster Linie* für die schnelle Bereitstellung wissenschaftlicher Ergebnisse, ihren schnellstmöglichen Austausch und *ihre Verknüpfung* untereinander gedacht. Die Zahl der seitdem gestarteten, auf Digitalisierung basierenden Forschungsprojekte wuchs vermutlich exponentiell – ebenso wie die Zahl jener Projekte darunter, deren Förderung inzwischen eingestellt wurde und deren Daten und Ergebnisse längst wieder im „information black hole“ (Cerf) verschwunden sind.¹

Vinton „Vint“ Cerf – mit Robert E. „Bob“ Kahn als Entwickler der Protokollfamilie TCP/IP zu Anfang der 1970er Jahre einer der „Gründerväter“ des Internet – warnt seit einigen Jahren davor, dass wir nicht nur Forschungsdaten, sondern *alle* digitalen Daten verlieren werden:

„We are nonchalantly throwing all of our data into what could become an information black hole without realising it. We digitise things because we think we will preserve them, but what we don't understand is that unless we take other steps, those digital versions may not be any better, and may even be worse, than the artefacts we digitised.“²

¹Beispielsweise Cerf 2015; zum WWW siehe Cerf 2014.

²Cerf 2015

Gemäß Cerf läuft unser digitales Zeitalter aus der Sicht zukünftiger Generationen Gefahr, ein „digital dark age“ zu werden, weil der größte Teil unserer Kommunikation und Forschung, eigentlich die gesamte Kultur im weitesten Sinne, inzwischen in digitalisierter Form über das Internet gespeichert und ausgetauscht wird – einer Struktur, die zwar für diese massenhafte Anwendung nie ausgelegt war, sie bis heute aber erstaunlich gut bewältigt.

Allerdings ist die gesamte heutige Hard- und Software in absehbarer Zeit obsolet – und es ist fraglich, ob und wie zukünftige Generationen Interesse, Zeit, Geld und Arbeitskraft aufbringen werden, unsere dann „alten“ – und aus ihrer Sicht vermutlich auch: veralteten – Daten zu retten.

Man sollte Cerfs Warnungen ernst nehmen, denn er ist jemand, der das Internet entscheidend mitprägte und seine Entwicklung in maßgeblicher Position begleitete. Sein eigener Vorschlag angesichts des drohenden Daten-Nirwanas ist die Schaffung einer Meta-Maschine, die er *Digital Vellum* nennt:³ Sie soll im Wesentlichen aus Hardware samt Betriebssystem bestehen, welche eine „virtuelle Maschine“, also einen „simulierten Computer“ bereitstellt, der die Daten *und* die zugehörige Software langfristig aufbewahren kann. Strukturell dürfte dies im Grunde der zur Zeit geplanten *Nationalen Forschungsdateninfrastruktur* entsprechen. Allerdings ist Cerfs *Digital Vellum* bereit einige Jahre in der Entwicklung und meines Wissens noch nicht fertig, geschweige denn: frei verfügbar. Sicherlich hätte ein Konzern wie Google, zu dessen Vizepräsidenten Cerf gehört, die Mittel, eine solche Entwicklung schnell voran zu treiben – aber ein profitorientiertes Unternehmen kann kaum daran interessiert sein, eine solche Lösung kostenlos (und werbefrei) als freie Software mit gegebenenfalls ebenso freiem Hardware-Design bereit zu stellen. Von Unternehmen, die ihr Geld gerade damit verdienen, ihren Kunden immer neue Versionen ihrer (Hard- und) Software zu verkaufen, kann dies erst recht nicht erwartet werden. Eine Abhängigkeit von solchen kommerziellen Interessen muss also – nicht zuletzt auch aufgrund der in der Regel vermutlich befristeten Existenzdauer solcher Firmen – ausgeschlossen werden: Ein in einer solchen virtuellen Maschine abgelegtes Word-Programm wird beim Start zum Öffnen eines 2020 erzeugten Dokuments im Jahre 2050 spätestens schon keinen Kontakt mehr zum Authentifizierungsserver herstellen können und also Start und Datenzugriff verweigern.

Man mag einwenden, dass ja der Inhalt des Dokuments in einem XML-Dialekt vorliegt und so prinzipiell rekonstruiert werden kann: Aber spätestens, wenn spezifische Layouts, Bilder oder Funktionen ebenfalls erhalten werden sollen, wird dies wohl kaum möglich sein. Zudem lassen sich in Textdokumenten zwar Forschungsergebnisse zusammengefasst darstellen, Forschungsdaten selbst aber schon kaum noch erfassen – zumindest nicht in einer sinnvollen digitalen Weise, die über eine Schreibmaschine hinausgeht.

Stattdessen wird es kaum noch ein Forschungsprojekt geben, das nicht eine mehr oder weniger aktuelle Datenbank-Management-Software nutzt. Zusätzlich wird zumeist eine spezifische Datenbank-Struktur (relational, objekt-orientiert, objekt-relational, graphenbasiert...) gemäß dem Forschungsdesign des Projekts entwickelt. Die Daten werden meist mit zusätzlicher Software beispielsweise für die Benutzeroberflächen erfasst und genutzt, die in sich ebenfalls rasant entwickelnden Skriptsprachen verfasst ist.

Versucht man sich zu vergegenwärtigen, *wie viele* Komponenten in der Hard-, vor allem aber in der Software am Zustandekommen einer jeweils projektspezifischen IT-Struktur beteiligt sind, erscheint es nahezu ausgeschlossen, diese vollständig und dauerhaft für sehr viele laufende

³Cerf 2015b.

und abgeschlossene Projekte konservieren zu können: Schon der Versuch, innerhalb eines einzelnen Projekts, erst recht vielleicht einige Jahre nach dessen Abschluss, eine Migration auf eine neuere Software-Version durchzuführen, ist in der Regel zumindest sehr kompliziert, häufig aber undurchführbar, beispielsweise weil keine ausreichende Dokumentation vorliegt und die ursprünglichen Entwickler und Wissenschaftler – oft in Personalunion –, deren Überlegungen und Forschungsinteressen das Design des gesamten Projekts prägten, gar nicht mehr verfügbar sind. Und selbst, *wenn* sie es heute sind und/oder eine tatsächlich vollständige Dokumentation hinterlassen haben, kann man kaum ernsthaft erwarten, dass beispielsweise in 100 Jahren oder später noch jemand Zeit, Geld und Arbeitskraft aufzuwenden bereit oder in der Lage sein wird, ihre Daten in einer dann nutzbaren Form – mitsamt der Software, der virtuellen Maschine und dem *digital vellum* oder der dazugehörigen Forschungsdateninfrastruktur – „wiederzubeleben“. Die unter anderem deshalb angestrebte, aber meines Wissens selten vollständig realisierbare „Abstraktion“ der Daten und ihrer Strukturierung von einer spezifischen Hard- und Software – beispielsweise in XML – soll dies zwar verhindern, aber zumindest in den *Digital Humanities* ist mir keine Projekte bekannt, denen es zum Beispiel gelungen wäre, Daten mittels des ISO-Standards *CIDOC-CRM* auszutauschen. Aber dies bestätigt nur meine Erfahrungen aus circa 35 Jahren im Umgang mit Computern und circa 25 Jahren Mitwirkung in Projekten der *Digital Humanities*.

Versteht man unter Langfristigkeit und Nachhaltigkeit *nicht* – wie gegenwärtig wohl noch diverse nationale Forschungsförderungsinstitutionen – eine Zeitspanne von 15-20, bestenfalls 50 Jahren (nur für TXT und PDF/A), hält man also digitale Forschungsdaten und -ergebnisse für wichtig genug, um sie genau so lange verfügbar zu halten, wie solche früherer Jahrhunderte auf Papier, wird man erst recht Zweifel an heutigen Lösungsansätzen hegen: Denn die tatsächlich *langfristige* Speicherung von Daten und Software verlangt nicht nur ihre einmalige Übertragung in ein *digital-vellum*-artiges System, sondern die *fortwährende* Migration auf die nächste oder spätestens übernächste Generation von Hard- und Software, und zwar nicht nur für die Projektdaten, sondern auch für das „Hostsystem“, beziehungsweise eben das *Digital Vellum* selbst.

Dies erfordert demnach, wie Cerfs Kollege Robert E. Kahn, verdeutlichte, auch eine *soziale Infrastruktur*, eine Institution, die von *langfristiger* Dauer und in der Lage ist, solche Systeme samt der erfassten Daten zu betreuen.⁴

Dem möchte ein Vorschlag Alan Kays für „digitale Tontafeln“ (*cuneiform tablets*) begegnen: Kay, der seit Ende der 1960er Jahre beim XEROX Palo Alto Research Center (PARC) die Entwicklung der Programmiersprache und -entwicklungsumgebung *Smalltalk* samt „virtueller Maschine“, die Konzeption und Realisierung graphischer Benutzeroberflächen sowie des objekt-orientierten Paradigmas entscheidend prägte, schlägt „selbsterklärende“ Datenträger – aktuell noch in der Form von CDs oder DVDs – vor, die es selbst dem interessierten „Archäologen“ noch in tausenden von Jahren erlauben sollen, in kürzester Zeit und ohne allzu großen Aufwand alle erforderlichen Strukturen zu rekonstruieren, die notwendig sind, um den Datenträger zu lesen und die darauf „eingefrorene“ virtuelle Maschine – nicht zufällig am Beispiel eines *Smalltalk*-Systems demonstriert – samt der darin erfassten Daten zu nutzen. Für die langfristige Speicherung nicht mehr zu verändernder Daten mag dieses Konzept geeignet erscheinen: Aber es geht damit eigentlich nur wenig über die „Speicherung“ auf bisherigen Schriftträgern wie Stein, Ton, Metall, Pergament oder Papier hinaus. Ein *wesentlicher* Vorteil der Digitalisierung ist aber gerade, dass

⁴Kahn 2016.

große Mengen erfasster Daten und zu ihrer Nutzung und Auswertung geschriebener Software nicht nur „fix(iert)“ genutzt, sondern auch kontinuierlich erweitert und angepasst werden können – möglichst ohne, dass die Informationen oder Metadaten der ursprünglichen Daten verloren gehen oder verfälscht werden können.

Zusammenfassend lässt sich sagen: Es bestehen folgende Probleme:

- aktuelle Nichtverfügbarkeit langfristig (= > 50 Jahre) stabiler Systeme;
- damit nicht garantierten Nutzbarkeit heutiger digitaler Daten;
- „Wildwuchs“ durch kommerzielle Interessen „geschützter“ Hard- und Software;
- zeitlich begrenzte Finanzierung digitaler Forschungsprojekte;
- damit ausgeschlossenen Möglichkeit, die *notwendige* regelmäßige Migration aller Daten und Software *langfristig* zu sichern;
- und die vernachlässigten Gefahr durch grundlegende IT-Paradigmenwechsel.

Angesichts all dessen erscheint es mir geradezu *ausgeschlossen*, dass der Versuch tatsächlich gelingen kann, diesem „Dschungel“ oder „Zoo“ an Hard- und idiosynkratischer Software in irgendeiner Form Dauerhaftigkeit zu verleihen, indem man das alles in ein weiteres, nicht minder komplexes Forschungsdatenmanagementsystem, ein Hostsystem wie das *Digital Vellum*, überträgt, zumal dessen langfristige Existenz von derjenigen einzelner Institutionen oder -verbände und vor allem einer verlässlichen Finanzierung abhängig wäre, die zumindest heutige Forschungsförderinstitutionen gar nicht leisten können und wollen.

Hinsichtlich der erwähnten, von Robert E. Kahn geforderten „social structure“, welche für die langfristige Datensicherheit unverzichtbar ist, darf zudem gelten, dass heutige Firmen und Institute diesbezüglich wohl nicht ausreichend sind; alte Universitäten, Archive oder Museen oder eben Bibliotheken kämen da vielleicht eher in Frage, zumal gerade Bibliotheken und Archive zudem nicht nur über die Kompetenz zur Speicherung von Daten, sondern auch zu deren Erschließung verfügen.

Lösungsvorschlag: ein radikaler Neubeginn

Aufgrund des bisher Geschilderten ist es meines Erachtens notwendig, eine vorzugsweise internationale Institution mit „Ewigkeitsgarantie“ zu schaffen, die vergleichbar einer Nationalbibliothek, einem Nationalarchiv oder -museum die dauerhafte Speicherung und Verfügbarkeit *aller* ihr übertragenen Daten garantieren kann. Dabei kann es sich jedoch nicht um eine einfache Sammlung gängiger, nahezu beliebig komplexer, projektspezifischer Hard- und Softwarelösungen handeln. Stattdessen sollte diese Institution der Entwicklung und Pflege (*Maintenance*) einer *vollständig neuen Hard- und Software-Infrastruktur* dienen, welche alle Erfahrungen der bisherigen Entwicklungen berücksichtigt, vor allem bekannte Fehler vermeidet, möglichst sicher und nachhaltig („sustainable“) konzipiert ist und nach den Regeln der *Free Software* nicht nur im Quellcode zugänglich (= open source), sondern auch für alle Interessierten und Zwecke frei

nutzbar sein sollte. Allerdings müsste diese Institution *immer* das Recht haben, die Rück- beziehungsweise Übernahme von Projektdaten und -software zu verweigern, wenn deren Anpassungen nicht zuvor mit ihr abgesprochen, auf Kompatibilität mit dem „Hauptsystem“ geprüft und dokumentiert worden sein sollten.

Das System selbst müsste zum einen auf einer freien, offenen, modularen Hardware-Architektur (oder einer Gruppe von Architektorentwürfen) beruhen, die von jeder interessierten Firma oder Institution implementiert werden kann: So könnten nicht nur kostspielige Monopolbildungen – Stichwort: *Wintel*-Allianz – vermieden, sondern auch notwendige Änderungen schneller umgesetzt werden. Dieses System müsste zudem den schnellen Austausch aller Komponenten ohne Gefährdung der Gesamtarchitektur erlauben. Mit dieser Hardware wäre die eigentliche – natürlich ebenfalls freie – Software (Betriebssystem, *Middleware* und Anwendungssoftware wie Office- und Datenbank-Programme aber auch Peripherie-Treiber, Programmiersprachen et cetera) nur durch eine möglichst kleine „virtuelle Maschine“ verbunden, die sich schnell und einfach an sehr viele oder neue Hardware-Strukturen anpassen ließe: *Smalltalk* läuft beispielsweise in einer solchen, sogar zur Laufzeit veränderbaren „virtuellen Maschine“ auf über 100 Prozessorarchitekturen.

Die für die eigentliche Projektarbeit notwendigen Software-Pakete sollten dann zum Beispiel nur je *eine*, jedoch umfassende Lösung für Office-Funktionen und Berechnungen sowie Datenbanken bereitstellen und gegebenenfalls neu zu definierende Standard-Datenformate verwenden. Die Datenbanksoftware müsste jeweils eine Lösung für die verschiedenen gebräuchlichen Paradigmen – relational, objekt-orientiert, graphenbasiert et cetera – enthalten. Das Ganze wäre ähnlich den heutigen *Docker Containern* in Linux zwischen verschiedenen Systemen weitgehend abstrahiert übertragbar.

Idealerweise sollten *alle* in so einem System vorhandenen Daten über eindeutige Identifier in der Art von „URLs“ adressierbar sein, wie es im Prinzip im System HTTP/WWW mit der an die Unix-Struktur angelehnten Trennung hierarchischer Struktur bereits angelegt ist. Damit wären die darin abgelegten Daten jederzeit und von überall erreichbar. Links zwischen diesen Daten wären ab einem gemeinsamen Wurzelverzeichnis stabil. Diese sollten aber endlich – wie in den ebenfalls schon fast 25 Jahre alten WikiWiki-Systemen automatisch ihre Backlinks erzeugen: Das heißt, beim Verlinken eines Datenobjekts X mittels eines Link-Objekts Y auf ein anderes Datenobjekt Z ($X \Rightarrow Y \Rightarrow Z$) würde in Z zugleich der Back-Link ($Z \Rightarrow Y' \Rightarrow X$) angelegt, so dass Kommentare, Zitationen und so weiter immer in beide Richtungen sichtbar wären: Die heute übliche unidirektionale Verlinkung bedeutet ja gegenüber der Fußnote auf dem Papier keinen Fortschritt, solange nicht im verlinkten Datensatz *von Hand* ein Back-Link angelegt wird.

Natürlich müssten alle Objekte beziehungsweise ganze Strukturbäume innerhalb des Gesamtsystems oder seiner in virtuellen Maschinen laufenden Untersysteme immer schützbar und dem Urheber eindeutig zuweisbar sein. Bei entsprechendem Design ließen sich so auch „Kataloge“ sämtlicher Daten jederzeit aktuell erstellen, weil das gesamte System im Prinzip zugleich seine eigene Datenbank wäre. Auch dies eine Idee, die natürlich nicht neu ist.

Damit könnte dann beispielsweise auch endlich das Problem einer echten *Open Access* Publishing sowie der *Open Data* für Forschungsdaten dauerhaft gelöst werden: Was nicht innerhalb des Systems auffindbar, erreichbar, verlinkbar *und frei nutzbar* wäre, würde mit der Zeit an Bedeutung verlieren. Durch die vollständige Offenheit der gesamten Plattform könnten zugleich

Bibliotheken und Universitäten oder andere Forschungseinrichtungen jeweils eigene, zueinander kompatible und über Netzwerke verbundene Repositorien anlegen. Und selbst die *Peer review* könnte ihrem hochtrabenden Namen endlich gerecht werden, indem *alle* Mitwirkenden untereinander tatsächlich als *Peers*, also Ebenbürtige, aufträten und *jede/r* jederzeit die Veröffentlichungen und Daten der anderen sehen, nutzen und kommentieren könnte, wie es Tim Berners-Lee für das WWW ursprünglich vorsah. Eine spezielle „Subplattform“ für „Verlage“ enthält das WWW dagegen nicht – sie hätte den Erfolg seines Projekts wohl auch aus kommerziellen Interessen verhindert.

Die immer wieder ins Feld geführte, angeblich nur durch Verlage zu garantierenden Qualitätskontrolle ist meines Erachtens ein Schein-Argument, denn während sie bisher angeblich durch Kleinstgruppen von Verlegern, Lektoren, Redakteuren und Gutachtern mit je eigenen – gerade im anonymen Review nicht immer offensichtlichen – Interessen gewährleistet sein soll, würde sie in diesem offenen System von Repositorien durch die prinzipiell *alle* Fachwissenschaftler umfassende *scientific community* tatsächlich realisiert. Die meines Erachtens zweifelhaften heutigen Bibliometrie-Systeme würden dadurch ebenso obsolet wie „Zitierseilschaften“ leicht erkennbar.

Die zu schaffende Institution sollte also nicht nur die strukturelle Form einer Bibliothek haben, sondern meiner Meinung nach auch an deren Arbeitsweise orientiert sein: Das heißt, sie würde das IT-System (wie bisherige Bibliotheken ihre Katalogsysteme) entwickeln und betreiben und an Kooperationspartner in Form einer virtuellen Maschine für deren eigene Projekte „ausliefern“. Die zu fordernde Kontrolle des Entwicklungsprozesse erlaubte die Garantie der (Rück-)Übernahme aller Daten nach Projektabschluss – wenn diese nicht ohnehin von vornherein auf einer von der Institution bereitgestellter *Cloud* „Software as a Service“ nutzen sollte. Regeln für den Zugriff auf beispielsweise sensible Daten oder die Erweiterung der Daten beziehungsweise ihre Übernahme in zukünftige Projekte ließen sich in solch einer einheitlichen Systemplattform nicht nur relativ leicht und standardisiert vereinbaren, sondern dank der modularen Struktur von gegebenenfalls ineinander „verschachtelbaren“ oder miteinander verlinkbaren virtuellen Maschinen auch leicht(er) realisieren.

Last but not least wäre eine solche institutionalisierte Infrastruktur nicht nur datentechnisch *nachhaltiger* als alles bisher Verfüg- und wohl auch Vorstellbare. Dank des offenen Designs von Hard- und Software ließen sich zweifellos auch „modulare“ Optimierungen schnell realisieren, die eine energetischen/ökologischen *und* technischen Nachhaltigkeit (beispielsweise durch Recycling von Bauelementen und geringstmöglichen Stromverbrauch) ermöglichten.

Man mag einwenden, dass ein solcher kompletter Neustart einer gesamten Forschungs-/IT-Infrastruktur *viel zu teuer* sei. Tatsächlich dürfte sie sich im dreistelligen Million- oder ein-, vielleicht sogar zweistelligen Milliardenbereich bewegen – insbesondere, wenn man beispielsweise die besten und entsprechend vergüteten Spezialisten engagieren will. . . Dies ist bisher wohl nur Google, Microsoft, Facebook, Apple oder der NSA möglich. Zumindest im letzteren Falle (beziehungsweise dem der Geheimdienste generell) könnte man sich ja aber auch einmal fragen, ob nicht die langfristige Rettung aller wissenschaftlichen Ergebnisse eine höhere gesellschaftliche Priorität haben sollte, als die Überwachung aller Bürger. Ganz abgesehen davon, dass auch die riesigen Investitionen der genannten und ähnlicher Firmen letztlich durch deren Kunden finanziert werden, die zugleich in der Regel auch Steuerzahler sind: Das Geld ist also prinzipiell längst vorhanden, nur falsch alloziert.

Ein weiterer Einwand könnte neben den Kosten für die Erstellung eines solchen Systems in den Kosten für die Migration aller heute vorhandenen Projektdaten von herkömmlichen Systemen auf dieses neue System gesehen werden. Aber solche Migrationsschritte werden in Zukunft *immer wieder* nötig und in absehbarer Zeit nicht mehr *möglich* sein, weil die verwendete Software nicht mehr weiterentwickelt oder neuesten Hardware- und Betriebssystemen nicht mehr unter vertretbarem Aufwand angepasst werden kann. Ich schätze, dass die *einmalige* Migration vorhandener, bewahrenswerter Daten und ihrer projektspezifischen Software auf das neue System *immer* weniger kosten wird, als deren Migration über die nächsten zwei oder drei Generationen herkömmlicher Systeme. Zwar wären natürlich auch innerhalb dieser neuen Gesamtplattform solche Migrationen notwendig, ihr Aufwand ließe sich aber auf ein Minimum reduzieren, weil *genau dies* bereits bei der Konzeption eine grundlegende Anforderung. Nach meinen Erfahrungen hat zumindest in den bisherigen institutionellen Strukturen *niemand* die Mittel und Möglichkeiten, oft nicht einmal ein Interesse, eine solche herkömmliche Migration auch nur *einmal* vorzunehmen – was dann zum „Friedhof der Projektleichen“ jeweils ein weiteres „Datengrab“ hinzufügt.

Man könnte zum Vergleich mit dieser scheinbar gigantischen Zukunftsinvestition auch andere kostspielige öffentliche Investitionen heranziehen, deren Nutzen für Staaten und ihre Bürger geringer, zweifelhaft oder sogar „negativ“ – also ein Schaden – ist. Stichworte hierzu wären Kriege und Rüstung, industrielle Landwirtschaft, Banken-, also Spekulantenrettung, Steuerbeitrag – allein *Cum-Ex* dürfte ein Vielfaches der einzuplanenden Projektmittel gekostet haben –, aktuelle Drohnen- und Mautdebakel und andere, absehbare Steuergeldverschwendung wie in allen bekannten ÖPP-Projekten.

Man sollte auch nicht vergessen, dass im Umfeld eines solchen offenen Systems eine Vielzahl von Dienstleistungen – von der Hardware-Entwicklung und Produktion bis zur Anpassung der Standard-Software an spezifische Projektbedürfnisse – entstehen und neben qualifizierten und langfristig sicheren Arbeitsplätzen auch wiederum Steuereinnahmen generieren würde.

Literatur

[Alle Internetquellen zuletzt abgerufen am 31.10.2019]

– Cerf 2014 = Cerf, Vinton: A Long Term View of the World Wide Web [Präsentation]: <https://vimeo.com/110794988>

– Cerf 2015a = Cerf, Vinton; Sample, Ian: [Interview] Google Boss warns of ‚forgotten century‘ with email and photos at risk, in: *The Guardian*, 13. Februar 2015: <https://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf>

– Cerf 2015b = Cerf, Vinton: Digital Vellum. [Präsentation]: <https://www.youtube.com/watch?v=STeLOogWqWk>

– Kahn 2016 = Kahn, Robert E.: Challenges and Opportunities for Digital Preservation (Keynote auf der Konferenz IPRES 2016 in Bern), Abstract unter: https://ead.nb.admin.ch/web/ipres2016/frontend/indexda54.html?page_id=1166

– Kay / Nguyen 2015 = Nguyen, Long Tien; Kay, Alan: The Cuneiform Tablets of 2015. Viewpoints Research Institute, VPRI Technical Report TR-2015-004, Los Angeles: 2015: http://www.vpri.org/pdf/tr2015004_cuneiform.pdf

Dr. phil. Bernd Kulawik, Bern/Berlin. Kontakt: be_kul@me.com