

Digitales Vergessen

Über Datenverlust, Datenformate und Datenarchivierung

Interview mit Reinhardt Altenhöner, Abteilungsleiter

Informationstechnik bei der deutschen Nationalbibliothek

Mit der steigenden Zahl digitaler Daten wächst die Notwendigkeit einer langfristigen zuverlässigen Archivierung, um zu verhindern, dass wichtige Informationen im Laufe der Zeit wegen Problemen bei der Speicherung oder aufgrund fehlender Zugriffsmöglichkeiten verloren gehen. Was heute noch aktuell ist, kann morgen schon Schnee von gestern sein. Oder haben Sie vielleicht Ihren Schallplattenspieler aufbewahrt? Was wird aus Ihren Videokassetten samt Videorekorder?

In der Vergangenheit gab es einige populäre Beispiele für die Schwierigkeiten der Archivierung digitaler Daten. So waren beispielsweise Daten, die von der Raumsonde „Pioneer“ 1979 vom Saturn übertragen und bei der NASA auf vier verschiedenen Datenträgern (Magnetbänder, Lochstreifen) gespeichert wurden, 1994 nicht mehr lesbar, da für diese Datenträger keine Lesegeräte vorhanden waren.

Die auf Magnetband gespeicherten Daten der US-Volkszählung von 1960 konnten nach Umstellung auf ein neues Speicherformat nur teilweise gerettet werden.

Die Beispiele scheinen in weiter Ferne und in anderen Größenordnungen zu liegen, aber die Anfrage eines Katasteramtes nach Digitalen Orthophotos (DOP) aus dem Jahre 1995 warf vor kurzem auch in der Landesvermessung und Geobasisinformation Brandenburg (LGB) die Frage auf, ob diese Bilder bereitgestellt werden können. Digitale Orthophotos werden für das Gebiet des Landes Brandenburg in

regelmäßigem Turnus aktuell hergestellt, ein Zugriff auf historische DOPs war bisher nicht notwendig. Die historischen DOPs sind auf Digital Audio Tapes im TIFF-Dateiformat im Archiv der LGB abgelegt. Die alten Lesegeräte standen noch bereit und das Bildformat war lesbar, so dass die „alten“ DOPs verfügbar waren.

Die Gründe für den Verlust digitaler Daten sind vielfältig und sollten jedem im täglichen Umgang mit digitalen Daten bewusst sein. Das Interview mit Reinhard Altenhöner, Abteilungsleiter Informationstechnik bei der Deutschen Nationalbibliothek, sensibilisiert für das Thema „Digitales Vergessen“:

Herr Altenhöner, Pessimisten orakeln, dass der langfristige Zugriff auf digitale Informationen eines der großen offenen Probleme unserer vernetzten Informationsgesellschaft ist. Teilen Sie diese Sorgen?

Ja, die teile ich – wobei - das sei gleich gesagt – ich nicht glaube, dass wir ernsthaft einen Totalverlust befürchten müssen.

Tatsächlich aber haben wir heute eine Situation, in der noch kaum bewusst ist, dass digitale Daten gefährdet sind – und in diesem mangelnden Bewusstsein bei vielen Datenproduzenten liegt eines der ganz wesentlichen Probleme.

Bewusst ist uns in der Regel, dass der physische Träger, auf dem die Daten zum Beispiel eine Sammlung von Dokumenten oder Bildern liegen, unbenutzbar werden könnte, beispielsweise durch eine physische Beschädigung. Oder aber ein geeignetes bzw. funktionierendes Lesegerät steht nicht (mehr) zur Verfügung. Wer hat etwa heute noch ein 5,25“ Laufwerk? Viel problematischer ist aber eigentlich, dass unser Wissen darüber, welche Eigenschaften und Anforderungen zum Beispiel ein bestimmtes Dateiformat zu einem bestimmten Zeitpunkt hatte, verloren zu gehen droht.

Welche Anforderung sind das? Wie können die Informationen darüber erhalten werden?

Jedes digitale Dokument benötigt eine definierte Hard- und Softwareumgebung: Das fängt häufig schon bei der Hardware an – eklatant zum Beispiel bei spezifischen Erweiterungen – und setzt sich bei Betriebssystemständen fort, speziellen Treibern, aber auch bestimmten Softwareständen und wird dann konkret sichtbar in den eigentlichen Viewern, die ich brauche, um die digitale Information anzuzeigen oder auch editierbar zu machen. Konkret: Schon heute kann es schwierig sein, auf ein Dokument aus den Achtziger Jahren zuzugreifen: Beispielsweise kann das Dokument mit Hilfe eines wenig verbreiteten Tools erstellt worden sein und dementsprechend in einem Dateiformat daherkommen, über das ich mich heute nicht ohne Weiteres

informieren kann. Die Dateikennung des digitalen Dokuments selbst reicht oft nicht aus, die notwendige Abspielumgebung herauszubekommen. Und bislang kümmert sich schlicht kaum jemand um diese Informationen. Wir von der Deutschen Nationalbibliothek haben seit 2006 den Auftrag, die so genannten „Netzpublikationen“ zu sammeln und zu erschließen, zu archivieren und auf Dauer zugänglich zu halten. Daher bemühen wir uns, schon zum Zeitpunkt der Überführung in unser Archiv möglichst viel an Information zur Entstehungsumgebung des Objekts und zu seinen spezifischen Eigenschaften zu erhalten, zu extrahieren und zu speichern – Wissen, das uns später helfen wird, die Datei lesbar zu machen oder Erhaltungsmaßnahmen einzuleiten.

Sie hatten die Dateiformatproblematik schon angesprochen. Wo sehen Sie hier konkret die Gefahren?

Die Informationstechnik hat in den vergangenen 40 Jahren zahllose Technologie- und Innovationssprünge durchlaufen – das spiegelt sich in der Vielzahl der entstandenen Programme und Werkzeuge wider. Nahezu jedes dieser Produkte erzeugt spezifische Objekte in einem für dieses Produkt definierten Dateiformat. Das bedeutet praktisch, dass wir ein erhebliches Mengenproblem haben, denn es sind zehntausende solcher Formate, mit denen spätere Nutzer potentiell umgehen müssen. Ein standardkonfigurierter PC „beherrscht“ heute einige hundert dieser Formate, aber wie stelle ich sicher, dass ich als Nutzer „historischer“ Objekte auch auf diese nicht mehr standardmäßig zur Verfügung stehenden Formate bzw. die dazu erforderlichen Werkzeuge zugreifen kann?

Wie beurteilen Sie verschiedene Datenformate auf ihre Eignung hinsichtlich der Langzeitarchivierung?

Bei den Dateiformaten sehen wir erhebliche Unterschiede, was ihre Eignung für die Langzeitarchivierung angeht. Dabei haben wir gar nicht so sehr im Auge, wie komplex ein solches Format sein kann. Vielmehr haben wir konkrete Anforderungen daran, wie gut und vollständig ein Format dokumentiert ist und ob diese Dokumentation und auch die Werkzeuge zur Anzeige frei verfügbar sind. Gerade diese Bedingungen erfüllt PDF/A als kürzlich international standardisiertes Format in bemerkenswertem Maße. Ganz generell ist es günstig, wenn ein Format textliche Information auch im Klartext enthält, hochgradig strukturiert ist und mit definierten externen Werkzeugen unabhängig von der Software, mit der es erstellt wurde, gelesen bzw. be- und verarbeitet werden kann.

Wie bewerten Sie die Ausgangssituation von Geoinformationen?

Die Geowissenschaften sind nach meiner Einschätzung eine hochgradig vernetzte Community, die schon früh erkannt hat, wie wichtig es ist, standardisiert vorzugehen und Werkzeuge und Datenformate transparent und interoperabel zu halten. Dass dies aber immer wieder großer Anstrengungen bedarf, zeigt die europäische INSPIRE-Initiative, die länderübergreifend bereits ein hohes Maß an Vereinheitlichung erzeugt hat.

Wie hoch dieser Aufwand bzw. wie weit der Weg ist, zeigt ja der Blick rückwärts: Ich muss als Nutzer solcher Daten immer wissen, welches methodische Framework zugrunde gelegt wurde, um zum Beispiel die Positionszuordnung vorzunehmen.

Wo liegen die Herausforderungen von digitalen Langzeitarchiven?

Spontan sehe ich da zunächst die große Menge der potentiell zu archivierenden digitalen Daten: Die Erzeugung digitaler Information erfolgt permanent und in rasant wachsenden Größenordnungen. Eine andere Herausforderung liegt natürlich in der Komplexität der Aufgabe selbst: Über die relativ kurze Geschichte der Informationstechnik hinweg sind bereits unzählige Formate und Formatversionen entstanden, die sich in Form digitaler Objekte als „digitales Erbe“ niederschlagen. Und ganz unabhängig von einer Bewertung ihrer Bewahrungswürdigkeit stellt sich zunächst einfach das Problem, für Konstellationen aus Softwareabhängigkeiten und Abspielumgebungen adäquate Verfahren zu definieren, mit denen die Zugreifbarkeit der Objekte auf Dauer gesichert werden kann.

Und da wir es mit einer sich ständig verändernden Welt zu tun haben – das Arbeitsgerät wird eben ständig weiterentwickelt –, besteht die Aufgabe kontinuierlich fort. Denn ein vor vier Jahren durchgeführter Migrationsschritt muss vielleicht heute neuerlich durchgeführt werden, weil die Formatversion, in die vor vier Jahren konvertiert wurde, wiederum in einer gängigen Anzeigenumgebung nicht mehr nutzbar ist. Dieser dauernde „Betreuungsbedarf“ stellt eine große Herausforderung dar.

Ein weiterer Aspekt ist die Frage, wie wir frühere Arbeits- und Rezeptionstechniken weitergeben wollen: einerseits haben wir es nämlich mit physisch vorhanden gewesenen historischen Eingabeeinheiten (beispielsweise ein Joystick oder spezielle Funktionstasten) zu tun, andererseits aber auch mit sich verändernden Methoden der

Anzeige von Daten und der Benutzung von Objekten. Dahinter stehen bestimmte Erfahrungen oder Gewohnheiten des Umgangs mit digitalen Medien zum Beispiel über die Bedienoberfläche. Ein common sense darüber, was man wo auf dem Bildschirm findet, mag heute einigermaßen etabliert sein, aber dieser Stand ist historisch gewachsen und verändert sich weiter. Das Problem stellt sich übrigens zum einen bei der Migration, also der Konversion der Objekte in neuere Formatumgebungen. Da unterschlagen wir quasi die alte Umgebung und müssen uns fragen lassen, wie authentisch das Objekt noch ist, Stichwort look & feel. Zum anderen ist auch die Emulation, also die Präsentation einer Datei in einer alten, virtuell auf einem Rechner bereit gestellten Umgebung, problematisch, denn da weiß ein heutiger Benutzer einfach nicht, wie er die Anzeigesoftware bedienen soll. Wie wir dieses Wissen erhalten und vermitteln wollen und wie die entsprechenden modernen Geräte alte Bedieneinheiten substituieren sollen, ist noch weitgehend offen.

Welche Möglichkeiten gibt es, archivierte Daten (alte Datenformate) für die Zukunft dauerhaft zu erhalten?

Abstrakt gesprochen gibt es zwei wesentliche Ebenen des Herangehens. Zunächst einmal müssen wir sicherstellen, dass die Daten genauso, wie sie einmal erzeugt wurden, auch wieder zur Verfügung stehen. Dieser Schritt, die ‚bitstream preservation‘ stellt die klassische Aufgabe von Rechen- und Datenzentren dar, die durch permanente Umkopier- und Checkprozesse sicherstellen, dass der Bit-Bestand der Daten als Ausgangspunkt zur Verfügung steht. Und zum Glück für uns „Bewahrer“ verändern sich die Grundgesetzmäßigkeiten

der IT z.B. bei der Codierung nur sehr langsam.

Die Frage aber, wie ich 30 Jahre nach Entstehung eines digitalen Objekts auf dieses zugreifen kann, ist damit noch nicht beantwortet. Darum gibt es Ansätze, die die Ausgangsobjekte schon beim Einspielen in ein Archiv ‚vorbehandeln‘, normalisieren bzw. in ein einheitliches, textbasiertes und strukturiertes Format überführen – und natürlich ist es einfacher, für ein einziges definiertes Format die geeignete Abspielumgebung dauerhaft anzubieten bzw. aktuell zu halten. Allerdings erkaufte man sich diesen Weg mit einem Verlust an Originalität und dem Risiko, dass es zu Informationsverlusten kommt.

Daher verfolgen wir und mit uns viele andere Archive einen anderen Weg: Wir gehen zunächst einmal vom Primat der Ausgangsinformation – also des ursprünglichen digitalen Objekts in seinem Dateiformat – aus. Zu diesen Objekten extrahieren wir automatisiert so viele Informationen wie möglich, mit deren Hilfe wir dann regelmäßig Migrationsschritte durchführen: das Objekt wird in einem vorab intensiv ausgetesteten und dokumentierten Arbeitsschritt in ein neueres Dateiformat überführt, das von aktuellen Abspielumgebungen verarbeitet werden kann. Es ist natürlich klar, dass dieser Schritt prinzipiell in bestimmten Rhythmen wieder anfällt, denn die Abspielumgebungen entwickeln sich stetig weiter. Und natürlich können bei aller guten Vorbereitung solche Konversionen / Migrationen auch fehlgehen oder Fehler eintragen – daher behalten wir zumindest das Ausgangsobjekt, um ggf. die Kette der Migrationen neu zu beginnen. Denn nach 30 Jahren sind vielleicht schon 4 - 5 solcher Konversionen erfolgt.

Das Wissen um die geeigneten Konversionswerkzeuge und auch die dafür notwendigen Informationen ist universell in dem Sinn, dass viele Archive weltweit vergleichbare Aufgaben haben und sich mit denselben Dateiformaten herumschlagen. Daher ist die internationale Kooperation und der Aufbau einer gemeinsam nutzbaren Infrastruktur, nicht nur technisch gemeint, so wichtig.

Manche Objekte eignen sich für solche Konversionen nicht bzw. können nicht ohne erhebliche Verluste konvertiert werden, wie zum Beispiel Multimedia-Objekte mit interaktiven Elementen oder auch ganze Anwendungen. Für diese Gruppe sehen wir den Weg der Emulation: Auf einer neueren Rechnerplattform wird eine historische Abspieulumgebung nachgebaut und wir können dem digitalen Objekt in seinem historischen Dateiformat ‚vorgaukeln‘, dass es auf ‚seiner Umgebung‘ abgespielt wird. Natürlich hat auch dieses Verfahren seine Grenzen, weil wie gesagt die historischen Bedienelemente, wie zum Beispiel Maus, bestimmte Funktionstasten oder ein Joystick, in der neuen Umgebung nicht mehr zur Verfügung stehen. Probleme gibt es übrigens häufig auch mit der Ablaufgeschwindigkeit.

Welche Überlegungen sollten hinsichtlich der Wahl des Datenformates bei zukünftig neu zu erfassenden Daten beachtet werden?

Entscheidend ist hier, dieses Format beherrschbar zu halten: Es sollte vollständig dokumentiert sein – eine Standardisierung ist hier ein guter Schritt – und offen nachnutzbar auch für andere sein. Hilfreich ist, wenn das Format Textinformationen menschenlesbar integriert, dabei sollte der eigentliche content bzw. Informati-

onsgehalt von Darstellungsinformationen getrennt sein. Ein weiteres Kriterium ist natürlich auch der Verbreitungsgrad, denn ein selten genutztes Format ist ein deutlich unsicherer Kandidat, als ein Format, das von vielen Menschen / Programmen genutzt wird.

Alle genannten Kriterien gelten in besonderem Maße für textbasierte Informationen; Bildinformationen sollten nach Möglichkeit in einem nicht komprimierten oder verlustfrei komprimierten Format abgelegt werden. Hilfreich ist es, wenn die Formate selbst schon Metainformationen zu den Objekten enthalten, die ggf. auch mit anderen Werkzeugen ausgelesen werden können.

Ein nicht direkt damit zusammenhängender Punkt betrifft Kopierschutzmechanismen bzw. Nutzungsbeschränkungen gleich welcher Art wie zum Beispiel auch Online-Aktivierungen / -überprüfungen: Formate, die originär solche Eingriffsmöglichkeiten anbieten, sind – sofern sie denn genutzt werden – grundsätzlich zunächst jedenfalls nicht langzeitarchivierungsfähig, da wir mit diesen Mechanismen/Informationen im Kontext von Migrationsprozessen nicht adäquat umgehen können.

Da drängt sich für mich gleich eine praktische Frage aus der Geodatenwelt auf. Nach all dem was Sie gerade zu den wünschenswerten Eigenschaften eines Dateiformates ausführten, scheint das XML-Format ein gutes Langzeitarchivierungsformat zu sein?

Ja, das kann ich bestätigen: Der Charakter dieses Formats, sozusagen im Klartext menschenlesbar Informationen zu transportieren, sein hoher, auch intellektuell nachvollziehbarer Strukturierungsgrad,

die klare Regelbasiertheit, die maschinelle Validierungsprozesse erleichtert sowie die hohe Flexibilität und die Breite der Einsatzgebiete lassen XML-basierte Formate im Vergleich sehr geeignet aussehen. Im Prinzip ist ja auch vorstellbar, aus der XML-Datei heraus das jeweilige Schema für die Darstellung mittels eines Viewer-Werkzeuges wiederherzustellen – auch wenn dies wegen der häufig hohen Komplexität sicher nicht einfach sein dürfte. Insofern muss man hier darauf achten, dass die notwendigen „Interpretationsregeln“ mit in den Archivierungsprozess eingehen. Und wie so oft ist eine gute Dokumentationslage und ein offener Zugang zu diesen Quellen von herausragender Bedeutung – dies belegt eindrucksvoll die aktuelle Diskussion rund um die Standardisierung von XML-basierten Dokumentenformaten.

Kommen wir von den Dateiformaten zu den Archivierungssystemen: Können Sie die Anforderungen an eine neue Generation von Systemen skizzieren, die den Ansprüchen einer digitalen Langzeitarchivierung genügt?

Es gibt schon heute eine Reihe durchaus konkurrierender Systeme, von denen man sagen kann, dass sie den Stand der Technik abbilden. Das bedeutet keinesfalls, dass der Stand der Technik „abgeschlossen“ sozusagen positiv erledigt ist, sondern wir bewegen uns auf einem Feld, in dem noch vieles im Fluss ist, viele Neuentwicklungen bzw. Funktionserweiterungen entstehen. Man kann vielleicht sagen, dass mittlerweile klar ist, was ein System zur digitalen Langzeitarchivierung leisten muss, Stichwort OAIS – die meisten Systeme basieren heute auf dem Standard „Open Archival Information System“ (ISO

14721:2003, aktuell im zyklischen Review-Verfahren). Viele der entstandenen Systeme sind aber Gesamtlösungen, die sich schlecht in existierende Umgebungen einbinden lassen, die wenig flexibel und schwierig zu erweitern sind. Es mangelt ihnen an Transparenz und Offenheit ihrer Schnittstellen und vor allem sind sie bislang kaum auf den Datenaustausch mit anderen Systemen ausgelegt, Stichwort Interoperabilität. Vor diesem Hintergrund ist es an der Zeit, über eine neue Generation von Systemen nachzudenken, die besser integrierbar, offenen Standards verpflichtet, transparent dokumentiert, flexibel anpassbar und erweiterbar ist. Und eine weitere wichtige Anforderung ist, sich von der Leitidee eines Systems zu entfernen und stattdessen modulare Teilsysteme zu spezifizieren. Das hat den besonderen Vorteil, dass so verstärkt auch Entwicklungen aus anderen technischen oder institutionellen Umfeldern aufgenommen werden können.

Zur Verwaltung digitaler Daten innerhalb eines Archivsystems werden – Sie wiesen bereits darauf hin - Metadaten benötigt. Neben den beschreibenden Metadaten für Recherchezwecke werden hier auch Metadaten zur technischen Bestandserhaltung benötigt. Welche technischen Metadaten sind das und wie können diese erfasst und archiviert werden?

Was hier geht, hängt sehr stark vom Anwendungsfall bzw. dem konkreten Format ab. Und neben den deskriptiven Informationen zum Objekt und den technischen Metadaten gibt es auch Strukturinformationen – also was gehört wohin oder wozu – und Verwaltungsinformationen sowie alles was zum rechtlichen Umfeld zu zählen ist.

Technische Metadaten können die Datei selbst technisch beschreiben, das Entstehungsumfeld bis hin zum Status einzelner beteiligter Komponenten bzw. Versionen, Hinweise zum Dateiformat enthalten und zu den Werkzeugen, mit denen die Daten genutzt werden.

Reinhard Altenhöner,
Deutsche Nationalbibliothek
r.altenhoener@d-nb.de

