WissGrid

Diskussionspapier

Arbeitspaket 3: Langzeitarchivierung von Forschungsdaten

Bitstream Preservation: Bewertungskriterien für Speicherdienste[1]

Autoren	Arbeitspaket 3: Langzeitarchivierung von Forschungsdaten
Editoren	Jens Ludwig, Torsten Rathmann, Harry Enke, Florian Schintke
Datum	16-02-2011
Dokument Status	Entwurf
Dokument Version	0.5

Abstract

Dieses Dokument gibt einen Überblick über Bewertungskriterien für vertrauenswürdige Speicherdienste, wie sie insbesondere für die Langzeitarchivierung von Forschungsdaten relevant sind. Es werden allgemeine, integritätsspezifische und sicherheitsspezifische Bewertungskriterien unerschieden, die eine Ausgangsbasis für Service Level Agreements (SLAs) bilden können.

Speicherdienstanbieter
: können anhand der Kriterien ihr Angebot für Nutzer genauer beschreiben.

Nutzer von Speicherdiensten
: können anhand der Kriterien ihre Anforderungen gegenüber Anbietern formulieren bzw. ihre Anforderungen mit einer Liste von Angeboten vergleichen und zielgerichteter einen passenden Anbieter wählen.

[1] This work is created by the WissGrid project. The project is funded by the German Federal Ministry of Education and Research (BMBF).

Inhaltsverzeichnis

1 Einleitung

Die Langzeitarchivierung digitaler Daten umfasst verschiedene Aufgabenbereiche, um die Nutzbarkeit von digitalen Daten langfristig sicherzustellen. Grundlegend ist, dass Daten hierfür auf der Speicherebene sicher vorgehalten werden müssen. Andere Aufgaben umfassen z.B. Bereiche wie Dateiformate, die veralten können, Kontextmetadaten, die zur Wiederverwendung benötigte Hintergrundinformationen angeben, oder organisatorische und finanzielle Fragen. Die Kompetenz und Verantwortung für diese Aufgaben verteilt sich oftmals über mehrere Akteure[2].

Im Folgenden werden nur Anforderungen und Bewertungskriterien für einen Teilaspekt der Langzeitarchivierung, der sogenannten *Bitstream Preservation*, definiert. Bitstream Preservation stellt sicher, dass Bits eines Datenobjekts sich nicht unbeabsichtigt verändern, zugreifbar bleiben und vor dem Ausfall, Verfall und Veralten der Speichermedien und -technologien bewahrt werden.

Warum Bitstream-Preservation?

Die Notwendigkeit zur Bitstream Preservation ist sicherlich jedem Nutzer digitaler Medien einsichtig, aber sie wird für Forschungsdaten auch von Forschungsförderern explizit verlangt.[3] Allerdings ist diese Aufgabe als nicht trivial einzustufen, auch wenn scheinbar Speicherplatz immer günstiger und zuverlässiger wird. Aber zu oft werden die vollen Betriebskosten eines Speicherdienstes vernachlässigt und theoretisch ermittelten Herstellerangaben wie MTTDL oder MTBF (Mean Time To Data Loss bzw. Mean Time Between Failure) wird zu viel Vertrauen geschenkt.[4] Als Beispiel sei nur das Problem der latenten oder versteckten Speicherfehler angeführt, dass Speichermedien nicht nur vollständig und unmittelbar offensichtlich versagen, sondern dass Fehler durch den gesamten Hard- und Software-Stack entstehen können und unter Umständen auch erst beim Auslesen entdeckt werden. Beispielsweise hat das CERN Anfang 2007 die Fehlerrate eines CERN Raid-Diskpools empirisch gemessen und in einem Datenbestand von ungefähr neun Terabyte 22 nicht abfangbare Checksummendifferenzen festgestellt.[5]

Warum Bewertungskriterien?

Im Folgenden werden Kriterien aufgeführt, die es Nutzern und Anbietern von Speicherdienstleistungen ermöglichen sollen, eine Bewertung unter dem Gesichtspunkt der Bitstream Preservation durchzuführen.[6] Ziel der Kriterien ist es nicht, mit ihnen absolute und unrealistische Sicherheitsgarantien zu formulieren, sondern Integritätsanforderungen und Kosten in ein sinnvolles Verhältnis zu

[2] Siehe Seite 6 – 8, 48ff der WissGrid Langzeitarchivierungsarchitektur, `http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.1-LZA-Architektur-v1.1.pdf`.

[3] Siehe Empfehlung Nummer 7 in: Deutsche Forschungsgemeinschaft, *Vorschläge zur Sicherung guter wissenschaftlicher Praxis*, Bonn 1998, `http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf` und *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*, Bonn 2009, `http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf`.

[4] Siehe z.B. Rosenthal, *Keeping Bits safe: how hard can it Be?*, Comm. of the ACM, Vol 53, No. 11, 2010.

[5] Panzer-Steindel, *Data integrity*, Cern 2007, `http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797`.

[6] Neben eigenen Erwägungen und den zitierten Quellen entstammen die Kriterien auch Baker et al., *A Fresh Look at the Reliability of Long-term Digital Storage*, EuroSys2006.

bringen und Vergleichbarkeit zu fördern.[7] Es wäre wünschenswert, dass Daten-/Rechenzentren ihre Speicherdienstangebote entsprechend der Kriterien beschreiben und die Nutzer bei der Bewertung unterstützen. Auch wenn nicht immer verbindliche Aussagen zu einzelnen Kriterien getroffen werden können, so ist eine rein informative Beschreibung des gegenwärtigen Stands bzw. des üblichen Erfüllungsgrades hilfreich. Das SLA4D-Grid-Projekt und das WissGrid-Projekt arbeiten derzeit daran, aus den Bewertungskriterien eine Vorlage für Service Level Agreements (SLAs) zu entwickeln, für die Fälle, in denen Kriterien vertraglich verbindlich definiert werden sollen.

Für Bitstream Preservation kann zwischen allgemeinen Speicherkriterien und den spezifischen, stärker auf Minimierung des Verlustrisikos zielenden Integritätskriterien unterschieden werden. In einer idealen Welt wäre (sofern es überhaupt Datenverlust gäbe) für den Nutzer nur ein Integritätskriterium wichtig, das die maximale Verlustwahrscheinlichkeit angibt, dass eine definierte Datenmenge in einem definierten Zeitraum verloren geht. Durch welche Technik die Integrität gesichert würde, ob sie durch viele Kopien oder durch besonders robuste Speichertechnologien erzielt würde, wäre für den Nutzer unerheblich. Da aber in der Praxis derzeit kein solcher Wert ermittelt werden kann, wird eine Sammlung von Integritätskriterien im Kapitel 3 vorgestellt. Eine Übersicht über allgemeine Kriterien wird in Kapitel 2 gegeben und Kriterien zur Vertraulichkeit von Daten finden sich im Kapitel 4.

Weitere Aspekte

Über die unten aufgeführten Kriterien hinaus sind weitere vertragliche Regelungen und Angaben für den vertrauenswürdigen Betrieb von Speicherdienstleistungen notwendig. Exemplarisch seien hier Angaben zu den Datenbesitzern/Verantwortlichen und Aktualisierungen des Angebots erwähnt. Ein Datenzentrum muss den Besitzer der Daten kontaktieren können, z.B. aus juristischen Gründen oder um über Zwischenfälle zu informieren. Auch ob und wie eine Weitergabe der Verantwortung erfolgt (z.B. aufgrund eines Arbeitsplatzwechsels des ursprünglichen Datenbesitzers), sollte geklärt sein.

Da sich die eingesetzten Technologien und die Rahmenbedingungen eines Datenzentrums schnell ändern können, ist davon auszugehen, dass eine anfängliche Beschreibung des Speicherdienstes während der Speicherdauer eines Datenbestandes ihre Gültigkeit verliert. Dies mag keine nennenswerte Relevanz haben, wenn es sich nur um Erweiterungen oder einfache Verbesserungen des Angebots handelt, es kann aber gerade bei Sicherheitstechnologien wichtig sein, Aktualisierungen vorzunehmen und alte Verfahren auslaufen zu lassen. Es sollte vorab geklärt sein, wie mit Änderungen und Aktualisierungen des Speicherdienstes umgegangen wird.

Hinweis zur Notation

Die Kriterien wurden nach ihrer Wichtigkeit für eine Einschätzung von Speicherdiensten für die Langzeitarchivierung im Bereich der öffentlich geförderten Forschung in die Kategorien „wichtig“ und „weniger wichtig“ eingeteilt. Die nach Auffassung der Autoren zentralen Kriterien sind mit dem Symbol ★ gekennzeichnet. Zu diesen Kriterien sollten Speicherdienstanbieter möglichst in jedem Fall Angaben machen.

[7] Siehe auch Wright et al., *The Significance of Storage in the „Cost of Risk“ of Digital Preservation*, IJDC 3, Vol 4, 2009. Ein Verfahren zur Evaluation von Alternativen zur Bitstream Preservation wird vorgestellt in: Zierau et al., *Evaluation of Bit Preservation Strategies*, iPRES 2010.

2 Allgemeine Service-Level-Kriterien für Speicherdienste

1. ★ **Datenvolumen:** Das bereitgestellte/benötigte Datenvolumen
   Einheiten: GB/TB/PB/...

2. ★ **Speicherdauer:** Der Zeitraum für den der Speicherdienst mindestens zur Verfügung gestellt wird/genutzt werden soll.
   Einheiten: Zeitdauer oder absoluter Zeitpunkt

3. ★ **Löschtermin:** Ist ein Löschtermin vorgegeben, müssen die betreffenden Daten an diesem Termin gelöscht werden. Maximale Aufbewahrungszeiträume oder Löschtermine sind in einigen Fällen durch den Gesetzgeber vorgeschrieben, insbesondere bei personenbezogenen Daten.
   Einheiten: Zeitdauer oder Zeitpunkt.

4. ★ **Sperrfrist:** In einigen Anwendungsfällen werden die Daten erst nach Ablauf einer Sperrfrist für die Öffentlichkeit bzw. die Fachöffentlichkeit zugänglich. Während der Sperrfrist können die Daten nur vom Produzenten bzw. von der Gruppe des Produzenten gelesen werden. Für Forschungsdaten räumen viele Datenzentren den Produzenten der Daten eine Sperrfrist ein, damit eine Erstverarbeitung in Ruhe möglich ist.
   Einheiten: Zeitdauer oder Zeitpunkt.

5. ★ **Zugriffsverfahren:** Über welche Arten von Zugang (Protokolle) kann/soll auf Daten/Dateien zugegriffen werden?
   Einheiten: Auflistung der unterstützten Zugriffsprotokolle, z.B. HTTP, SCP, GridFTP...

6. **Datenanforderungen:** Welche Größen können Daten/Dateien haben? Gibt es eine Limitierung der Anzahl? (z.B. um ein Bandspeichersystem nicht durch zuviele kleine Dateien zu belasten)
   Einheiten: max MB/GB/TB/PB/..., min MB/GB/TB/PB/..., max Anzahl.

7. **Zugriffsgrößen:** Wie lange dauert es von der Anfrage nach einer Speicheroperation bis zu ihrem Beginn maximal bzw. im Mittel? Wie viele Zugriffe auf archivierte Daten sind erlaubt/-geplant? Wie groß ist die maximal abrufbare/abgerufene Datenmenge in Summe, pro Zugriff und pro Datei?
   Einheiten: Maximalzahl Zugriffe/Zeiteinheit, maximale Datenmenge/Zeiteinheit. (Ggf. sind diese Einheiten auf bestimmte Zeiträume bezogen, z.B. wenn in weniger oder stärker nutzungsintensiven Zeiträumen, intensivere oder geringere Anforderungen erfüllt werden können.)

8. **Transfergeschwindigkeiten:** Wie schnell können Daten gelesen und geschrieben werden?
   Einheiten: min/max Datenmenge pro Zeiteinheit

9. **Verfügbarkeit:** Wie häufig ist das System nicht nutzbar?
   Einheiten: „9s“, Prozentangabe

10. **Maintenance/Service:** Wie ist die Reaktionszeit bei Problemen?
    Einheiten: Zeitdauer nach Problemeingang, ggf. auf verschiedene Zeiträume (z.B. Wochenende vs. Wochentag) bezogen.

3 Spezifische Bitstream-Preservation-/Integritätskriterien

Die folgenden Angaben beziehen sich jeweils nur auf den dauerhaften, langfristigen Speicherort. Eigenschaften von Kopien in temporären Zwischenspeichern (z.B. Plattencache/SAN vor einer Tape-Library) können für die Bitstream Preservation vernachlässigt werden.

11. ★ **Kopienanzahl:** Werden Kopien der Daten dauerhaft gespeichert? Je mehr Kopien dauerhaft gespeichert werden, desto unwahrscheinlicher ist es, dass alle Kopien defekt sind. In der derzeitigen Praxis ist es fast nur die Kopienanzahl, die bei unterschiedlichen Angeboten eines Dienstleisters tatsächlich variiert wird.
    Einheiten: Anzahl identischer Kopien

12. **Kopienunabhängigkeit:** Wie unabhängig sind die Kopien unter verschiedenen Gesichtspunkten? Erst die Unabhängigkeit der Kopien voneinander sichert, dass die Kopien nicht vom selben Fehler oder Ereignis betroffen sind.

    12.1 **geographisch:** Die geographische Unabhängigkeit ist z.B. für lokale Unfälle, Unglücksfälle, Naturereignisse, etc. relevant und kann nicht nur als Distanz, sondern auch durch Kategorisierung wie unterschiedliche Räume, unterschiedliche Gebäude/Gefahrenzonen bis hin zu unterschiedlichen Regionen/Ländern/Kontinenten ausgedrückt werden.
    Einheiten: z.B. unterschiedliche Räume, Gebäude, Gefahrenzonen und Entfernung in km

    12.2 **organisatorisch:** Unter organisatorischer Unabhängigkeit ist die Unabhängigkeit von den selben Personen, Arbeitsabläufen, Institutionen oder auch rechtlich-politischen und finanziellen Rahmenbedingungen zu verstehen.
    Einheiten: z.B. unterschiedliche Administratoren, unterschiedliche Institutionen/Unterauftragnehmer und ggf. unterschiedliche Gesetze/Länder

    12.3 **technologisch:** Viele Fehler entstehen durch Probleme mit einem bestimmten Typ, einer Version oder Produktionsreihe der verwendeten Speichermedien, Hardware oder Software. Das Risiko dieser Art von Fehlern lässt sich durch eine kontrollierte Heterogenität der eingesetzten Technologien vermindern.
    Einheiten: jeweils [Produktionscharge, Modell, Hersteller, Technologie] für Speichermedien, -Hardware und -Software

13. **Fehlerhäufigkeit der Kopien:** Angaben zu der Fehlerhäufigkeit von Kopien sollte idealerweise auf empirisch gemessenen Daten beruhen und nicht nur auf theoretischen Hochrechnungen wie z.B. den Herstellerangaben zur Speichermedienhaltbarkeit beruhen.

    13.1 **Ausfallrate Speichermedien:** unmittelbar sichtbares Versagen eines kompletten Datenträgers
    Einheit: möglichst die emp. Austauschrate (annual replacement rate (ARR)), falls nicht möglich, dann Hersteller MTTF/MTBF (mean time to/between failure) pro Speicherkapazität in MB/GB/TB/PB

    13.2 **Bitfehlerrate:** die Häufigkeit latenter/verborgener Fehler auf Bit- oder Blockebene,
    Einheit: Mean time to latent fault

    13.3 **Gegenmaßnahmen:** Maßnahmen um die Fehlerhäufigkeit zu reduzieren, Lesbarkeit zu gewährleisten und Alterungsprozessen der Datenspeicherung entgegenzuwirken, z.B. Häufigkeit des Umkopierens der Daten auf andere Medien.
    Einheiten: Freitext

14. **Integritätstests:** Integritätstest sind ein Mittel, um latente/verborgene Fehler zu entdecken.

    14.1 **Häufigkeit:** Je häufiger ein Integritätstest durchgeführt wird, desto schneller können Fehler entdeckt und die Integrität wieder hergestellt werden.
    Einheit: Dauer zwischen zwei Tests

    14.2 **Verfahren:** Unterschiedliche Integritätstests sind unterschiedlich zuverlässig, z.B. CRC vs. MD5 Checksummen.
    Einheit: Verfahren, Checksummenalgorithmus

    14.3 **Integrität der Checksummenkopien:** Prinzipiell sind auch Checksummen Datenobjekte, die verloren gehen können und nicht nur als einfache Kopie vorliegen sollten.
    Einheit: Verfahren

15. **Integritätswiederherstellung:** Idealerweise stellt ein Speicherdienst im Fehlerfall die Integrität wieder her, allerdings besteht eine gewisse Gefahr, dass Fehler repliziert werden.

    15.1 **Verfahren:** Existiert ein Mechanismus zur Integritätswiederherstellung?
    Einheiten: Ja/Nein, eingesetztes Verfahren

    15.2 **Dauer:** Je schneller die Integrität wiederhergestellt wird, desto kürzer herrscht eine erhöhte Verlustwahrscheinlichkeit, allerdings kann ein gleichzeitiger Produktivbetrieb verlangsamt werden.
    Einheiten: durchschnittliches Datenvolumen pro Zeiteinheit

4 Sicherheit und Vertraulichkeit

16. **AAI-Verfahren:** Beschränkung des Datenzugriffs im Rahmen von AAI-Verfahren

    16.1 **Authentifizierungsverfahren:** Nach welchem Verfahren wird sichergestellt, dass nur Berechtigte Zugang haben?
    Vokabular: Username/Passwort, OpenID, Zertifikat

    16.2 **Autorisierungsverfahren:** Nach welchem Verfahren werden Zugriffsrechte vergeben?
    Vokabular: XACML, Unix, NTFS

17. **Beschränkung des externen Datenzugriffs:** Auf welche Art wird der Zugang zu den Daten generell sicherheitstechnisch eingeschränkt? Beispielsweise, um sich gegen ein Aushebeln der AA-Infrastruktur zu schützen. Andere mögliche Maßnahmen können sein: separater virtueller Server je Kunde, IP-adressbasierter Zugriff (Whitelist, Firewall), Sichtbarkeit nur der jeweils vom authentifizierten Nutzer zugreifbaren Daten oder die Verschlüsselung von Daten.
    Einheit: Freitext

18. **Beschränkung des internen Datenzugriffs:** Auf welche Weise werden Zugriffe des Speicherdienstleisters auf die Daten eingeschränkt? Wie viele Personen haben beim Speicherdienstanbieter Zugriff auf die Daten? Wie viele Personen haben physikalisch Zugriff auf die Speichergeräte? Ist z.B. beim Zugriff auf die Daten beim Speicheranbieter ein Vier-Augen-Kontroll-Prinzip realisiert? Werden Daten verschlüsselt?
    Einheit: Freitext

19. **Schutz der Kopien vor äußeren Einflüssen:** Werden gesonderte Sicherungsmaßnahmen ergriffen, die die Zuverlässigkeit der Datenspeicher erhöhen, wie zum Beispiel konstante Luftfeuchtigkeit und Raumtemperatur, Absicherung gegen Umweltkatastrophen, Feuer, Diebstahl, . . . ?
    Einheit: Freitext

5 Checkliste

Kriterium	Anbieter	Kunde
Allgemeine Service-Level-Kriterien		
Datenvolumen ★		
Speicherdauer ★		
Löschtermin ★		
Sperrfrist ★		
Zugriffsverfahren ★		
Datenanforderungen		
Zugriffsgrößen		
Transfergeschwindigkeiten		
Verfügbarkeit		
Maintenance/Service		
Spezifische Bitstream-Preservation-/Integritätskriterien		
Kopienanzahl ★		
Kopienunabhängigkeit		
geographisch		
organisatorisch		
technologisch		
Fehlerhäufigkeit der Kopien		
Ausfallrate Speichermedien		
Bitfehlerrate		
Gegenmaßnahmen		
Integritätstests		
Häufigkeit		
Verfahren		
Integr. Checksummenkopien		
Integritätswiederherstellung		
Verfahren		
Dauer		
Sicherheit und Vertraulichkeit		
AAI-Verfahren		
Authentifizierung		
Autorisierung		
externer Datenzugriff		
interner Datenzugriff		
äußerer Einflüsse		