KOST Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen

Ein Gemeinschaftsunternehmen von Schweizer Archiven

Kolloquium Archivtaugliche Speicherinfrastruktur

Abschlussveranstaltung

Bern, 18. Dezember 2007

Programm

1. Begrüssung und Einführung (10')
2. Berichterstattung über die thematischen Veranstaltungen (60')
   a. Speicherplatz mieten
   b. Speicherzentrum
   c. Blackbox
   d. Datenauslagerung
   e. Speichern im Netzwerk

   Kaffeepause (14:40)
3. Analyse an Hand der gliedernden Gesichtspunkte aus der Einführungsveranstaltung (40')
   a. Archivische Anforderungen
   b. Technische Aspekte
   c. Organisatorisch-rechtliche Aspekte
   d. Kosten
4. Fazit und Abschlussdiskussion (40')

   Apéro (16:30)

Speicherplatz mieten
Inhalt

- Das Archiv mietet den benötigten Speicherplatz bei einem (kommerziellen) Anbieter für ~5 Jahre.
- Die Anforderungen werden in einem *Service Level Agreement* (SLA) festgehalten.
- Der Anbieter ist für die Datenpersistenz verantwortlich.
- Die Angebote sind sehr unterschiedlich gestaltet, müssen auf Kunden massgeschneidert werden.
- Referenten:
  COLT Telecom (Leistungserbringer),
  FAST LTA (Produktanbieter).
  Auch kantonale Informatikdienste sind als Anbieter denkbar.

Speicherplatz mieten
Erkenntnisse

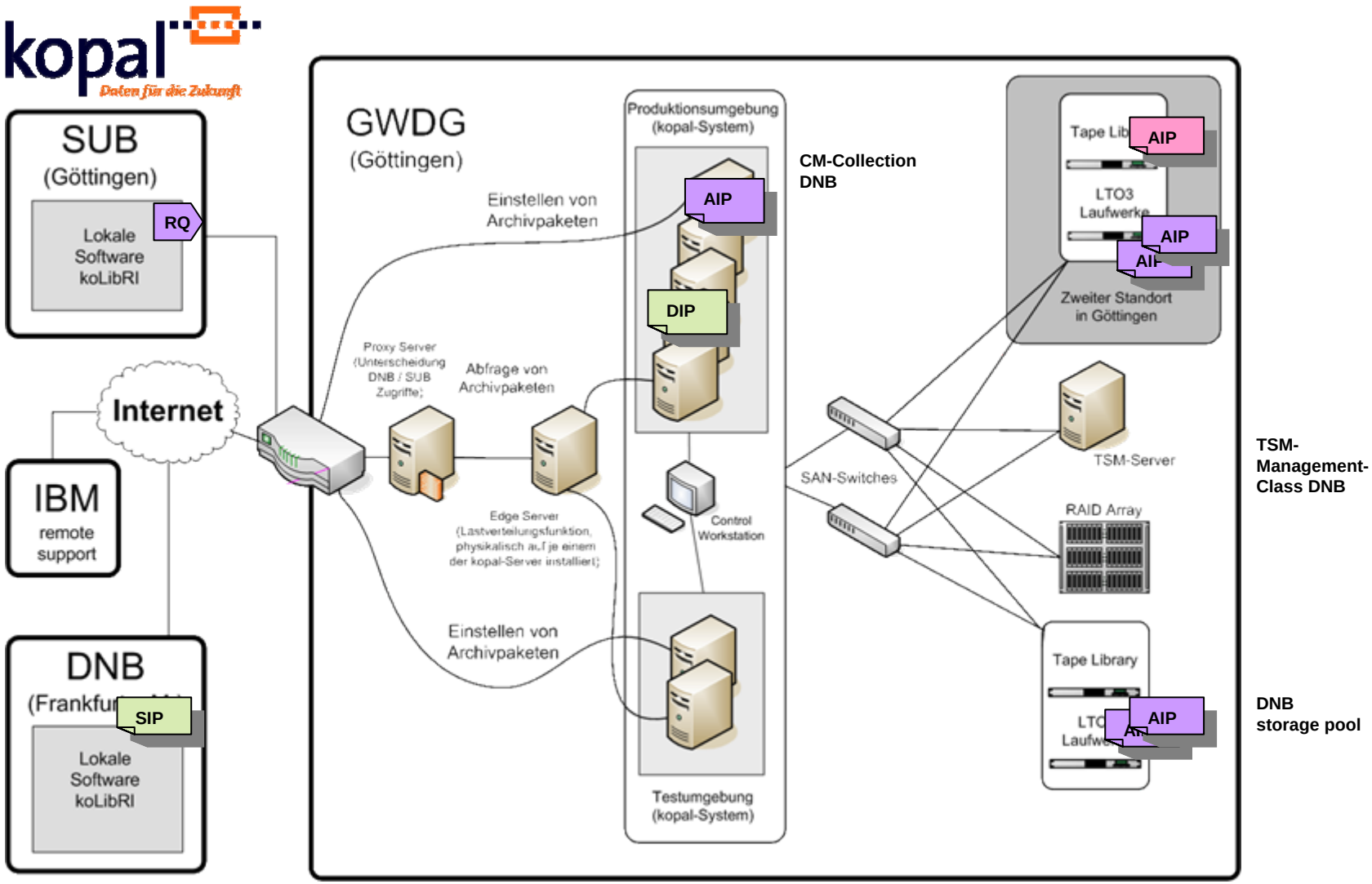
- Angebote ab Stange sind für Archive grundsätzlich wenig geeignet: Das Anforderungsprofil der Archive ist speziell.
- Die Formulierung des SLA ist deshalb entscheidend.
- Nicht zu viel bezahlen!
- Richtpreis FAST LTA: 3 SFr./GB/Jahr
- Die in Archiven zu erwartenden Datenmengen sind für kommerzielle Anbieter relativ gering, was den Preis hochtreiben kann. Kooperationen und gemeinsames Verhandeln werden empfohlen.

Speicherzentrum
Inhalt

- Ein Speicherzentrum ist ein Rechenzentrum, das primär Dienstleistungen im Bereich der Daten-speicherung und Langzeitarchivierung erbringt.
- Zwei Rechenzentren aus dem öffentlich-rechtlichen Bereich stellen ihre Erfahrungen beim Aufbau und Betrieb dieser Dienstleistung vor:
- *Dagmar Ullrich* von der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

- *Jürg Gemeinder* vom Verwaltungsrechenzentrum St.Gallen (VRSG)

Speicherzentrum *Workflow GWDG*



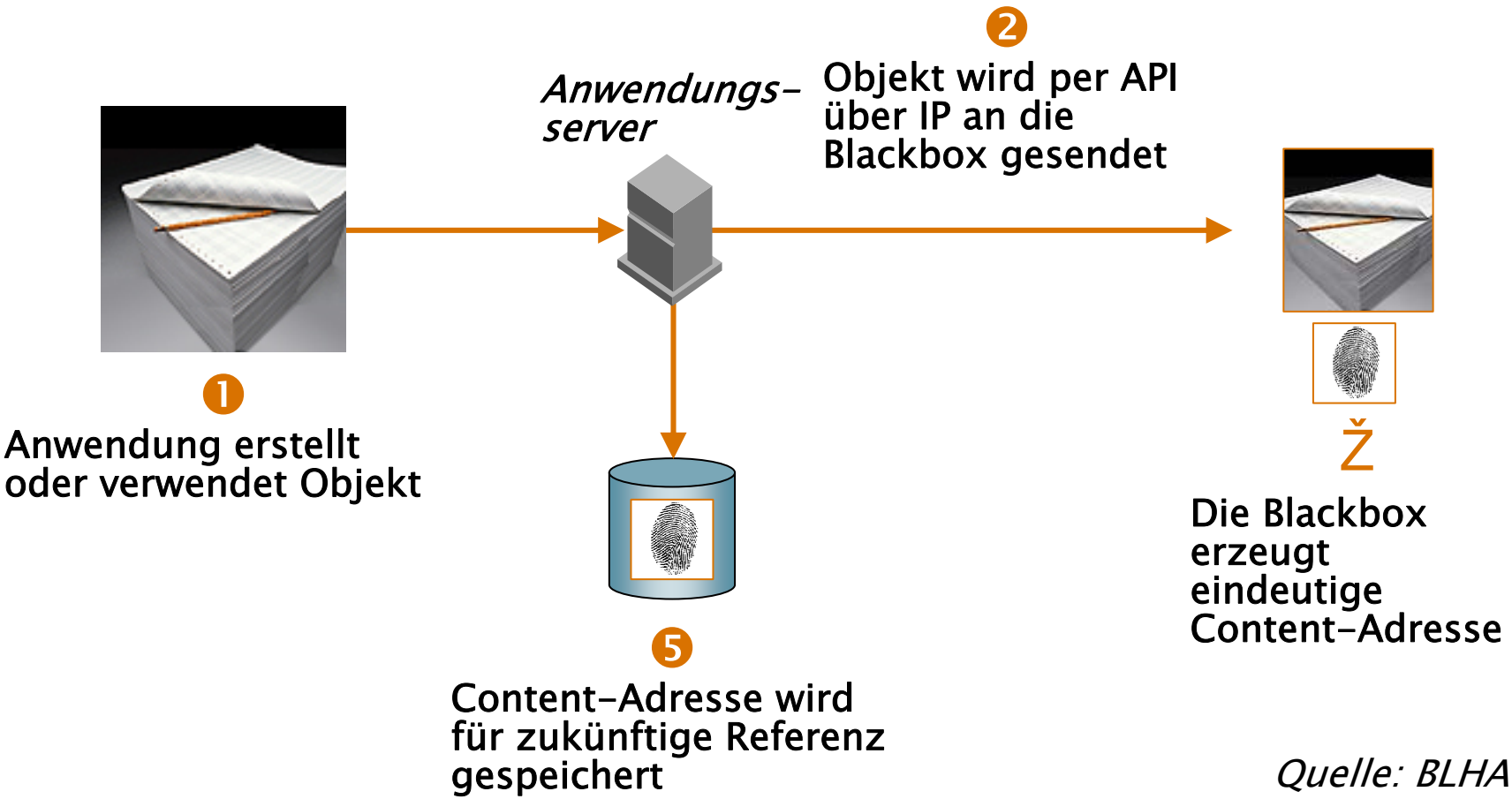
Speicherzentrum
Erkenntnisse

- Der Aufbau eines Speicherzentrums setzt ein minimales (Daten-) Volumen voraus und ist keine leichte Sache.
- Vom Neuaufbau eines Rechenzentrums raten die Referenten ab.
- Mieten bestehender Rechenzentrumsinfrastruktur scheint sinnvoller.
- Das Beherrschen der Hardwaremigrationszyklen ist eine der grössten Herausforderungen.
- Klare Trennung von Speichermanagement und Archivsoftware ist unerlässlich

Blackbox
Inhalt

- Das Archiv beschafft eine eigene Speicherbox.
- Die Blackbox stellt die Datenpersistenz sicher.
- Das Archiv übt die physische Kontrolle über die Box aus, weiss aber nicht, wie diese intern funktioniert.
- Der Wartungsaufwand ist gering und erfordert kaum Spezialistenwissen.
- Investitionen fallen schubweise an.
- Referenten:
  Brandenburgisches Landeshauptarchiv (betreibt Centera-Speicherlösung),
  IBM (Blackbox-Anbieter)

Blackbox
Workflow



Quelle: BLHA

Blackbox
Erkenntnisse

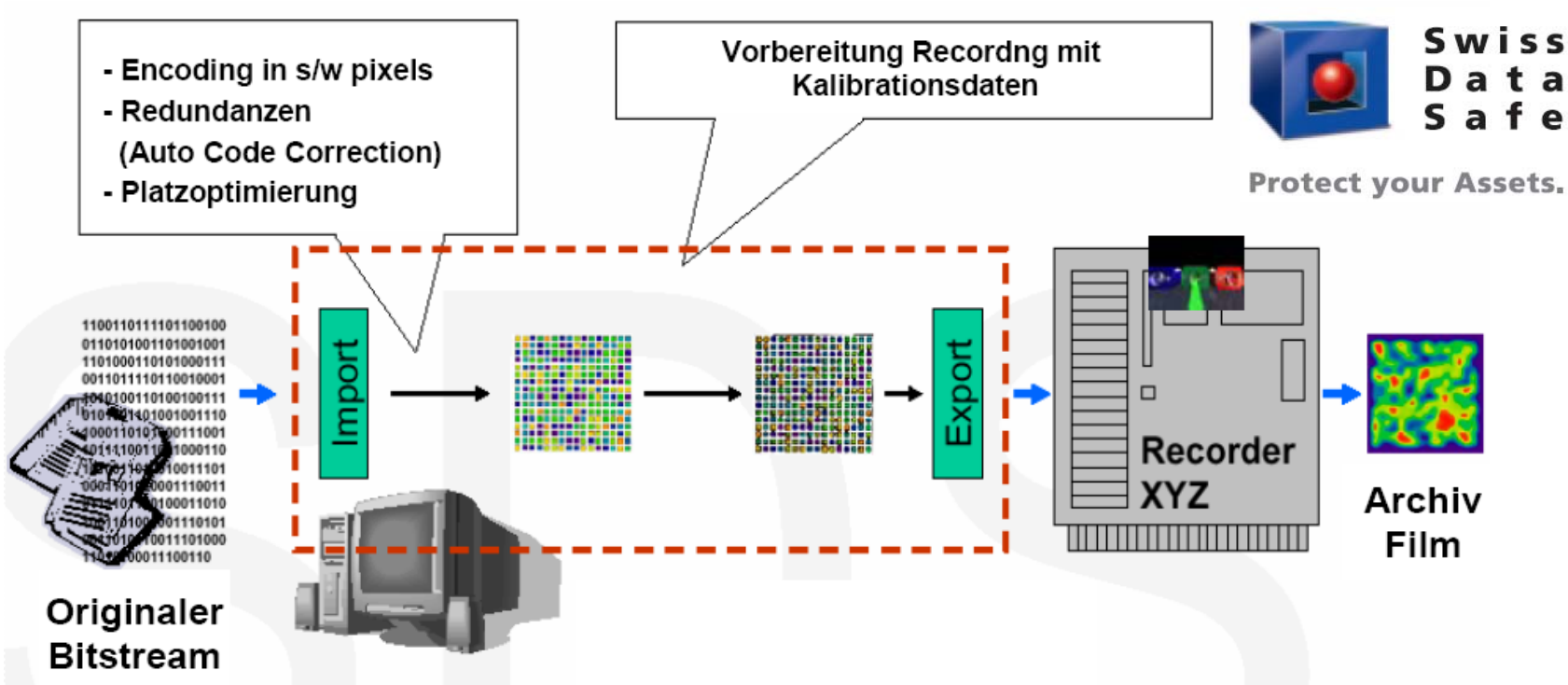
- Das Anforderungsprofil eines Archivs spricht eher für eine Tape- als für eine Disklösung.
- Die Blackbox gewährleistet eine hohe Datensicherheit, ist aber tendenziell eher teuer (Richtpreis BLHA: 6.85 SFr./GB/Jahr).
- Die Blackbox wird über ein proprietäres API angesprochen. Das ist nicht ideal, aber kein Problem.
- Für die in den Staatsarchiven in den nächsten Jahren anfallenden Datenmengen ist eine Tape-Library zu gross; ein einfaches Disk-System wäre grundsätzlich besser geeignet.

Datenauslagerung
Inhalt

- Die digitalen Daten werden auf ein langzeit-stabiles Speichermedium umkopiert und in dieser Form ausgelagert.
- Die zwei Referenten präsentieren die Lösung aus Sicht der Forschung und in der kommerziellen Anwendung:

- *René Meier,* Swiss Data Safe AG
- *Peter Fornaro,* Peviar, Imaging & Media Lab (IML)

Datenauslagerung
Workflow



Datenauslagerung
Erkenntnisse

- Die Idee der Auslagerung der Daten aus der operativen Informatikinfrastruktur gewinnt mit dem Einsatz von Mikrofilm als Trägermedium eine neue Dimension (> 100 Jahre).
- Theoretisch können auf einen 35-mm Film von 600m Länge (entspricht etwa einer Rolle von 40 cm Durchmesser) ungefähr 250 GB Nutzdaten geschrieben werden.
- Das Lesen ist mit einem handelsüblichen Scanner und einer Opensource Software möglich.
- Eine Standardisierung hat noch nicht stattgefunden.

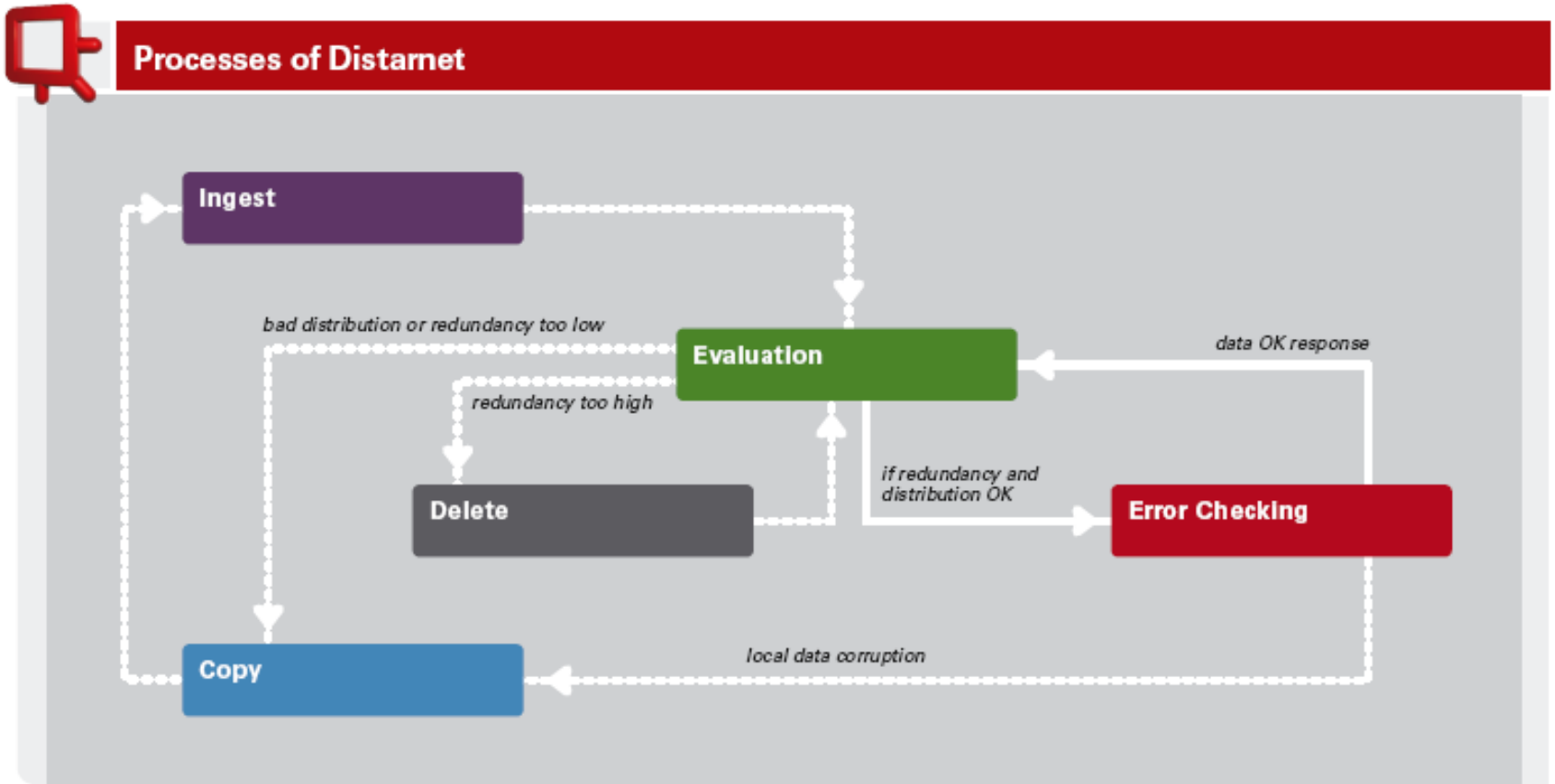
Speichern im Netzwerk
Inhalt

- Die digitalen Daten werden in einem Peer-to-Peer-Netzwerk von miteinander verbundenen Servern gespeichert. Das Fehlen einer zentralen Instanz erhöht die Ausfallsicherheit.
- Ein Protokoll sorgt für Redundanz und Daten-integrität.
- Angestrebt wird einfache Administration.
- Die Kosten sind schwierig zu beziffern, da brauchbare Erfahrungen fehlen.
- Referate:
  distarnet (Forschungsprojekt Uni Basel),
  LOCKSS (verteilte Speicherlösung im Bereich e-journals).

dist.ar.net
Langzeitarchivierung digitaler Daten

LOCKSS

Speichern im Netzwerk
Workflow



Speichern im Netzwerk
Erkenntnisse

- Das Konzept ist aus archivischer Sicht interessant: Grosse intellektuelle Kontrolle über Speicherlogik, geografische Redundanz ist eingebaut.
- Zwei Varianten sind denkbar: Verbund mehrerer Archive oder archivinternes Netzwerk.
- Die Lösung kommt jedoch im Moment nicht in Frage, da noch nicht ausgereift. Es existiert auch keine archivische Anwendung.
- Weitere Forschung ist notwendig und sollte auch herkömmliche P2P-Systeme berücksichtigen.

Archivische Anforderungen (1)

- Archive haben ein spezielles Anforderungsprofil:
  - Verfügbarkeit kann niedrig sein (Nächte, Wochenenden).
  - Zugriffszeit muss nicht optimiert werden: Zugriffszeiten im Minutenbereich sind absolut akzeptabel.
  - Es gibt keine Transaktionen auf den archivierten Daten, nur inkrementelles Backup ist notwendig.
  - Der Zugriff auf die Daten erfolgt arbiträr.
  - Datenintegrität und Datensicherheit sind absolut zentral.
- Es existieren wenige Speicherplatz-Angebote, die auf diese Anforderungen zugeschnitten sind: Die Gefahr besteht, zu viel zu bezahlen oder zu wenig zu bekommen.
- Diese Probleme bestehen vor allem bei „Speicherplatz mieten“ und bei „Blackbox“.

Archivische Anforderungen (2)

- Die Preisunterschiede zwischen verschiedenen Angeboten erklären sich aus den unterschiedlichen Anforderungsprofilen, die sie erfüllen. (Siehe dazu auch die Zauberformel der Informationssicherheit: CIA, kurz für *confidentiality – integrity – availability* [Vertraulichkeit – Integrität – Verfügbarkeit].)

Das archivische Anforderungsprofil ist das wichtigste Werkzeug bei der Analyse und Beschaffung von Speicherlösungen.

Technische Aspekte
Innovation

Betrachten wir die Lösungen aus Sicht des Innovationspotentials und der vermutlichen Entwicklungsmöglichkeit:

- *Speichern im Netzwerk* ist wahrscheinlich die innovativste Lösung und hat in der Form der „Filesharing"-Lösungen im Internet schon sein Potential gezeigt.
- *Datenauslagerung* kann in standardisierter Form langfristig die Archive entlasten.
- *Speicherplatz mieten* und *Blackbox* Lösungen sind die aktuellen Angebote der Informatik-industrie.

Technische Aspekte
Best Practice

Betrachten wir die Lösungen aus Sicht der technischen Realisierbarkeit zum jetzigen Zeitpunkt und unter dem Aspekt „Best Practice“:

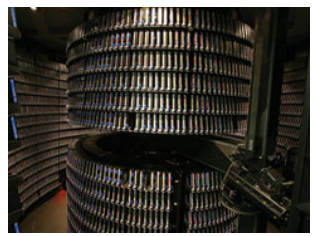
- Eine *Blackbox* Lösung kann heute als ausgereift betrachtet werden und lässt dem Archiv viel Gestaltungsfreiraum.
- *Speicherplatz* mieten verursacht wenig Aufwand im Archiv, schafft aber externe Abhängigkeiten.
- Das *Speicherzentrum* (Rechenzentrum) ist die klassische Lösung ohne besonderes Entwicklungspotential, aber mit viel Erfahrungshintergrund

Technische Aspekte *Speichermedien*

Betrachten wir die Lösungen aus Sicht möglicher Speichermedien ergibt sich folgendes Bild:

EMC CLARiiON

- Für Datenbestände von 1 bis 20 TB, wie wir sie für die nächsten fünf Jahre bei Staatsarchiven erwarten, sind *„spinning Disk“* das Medium der Wahl.



SDSC Tim Mcnew

- Für Datenbestände >20 TB muss eindeutig *„Tape Library“* gewählt werden.

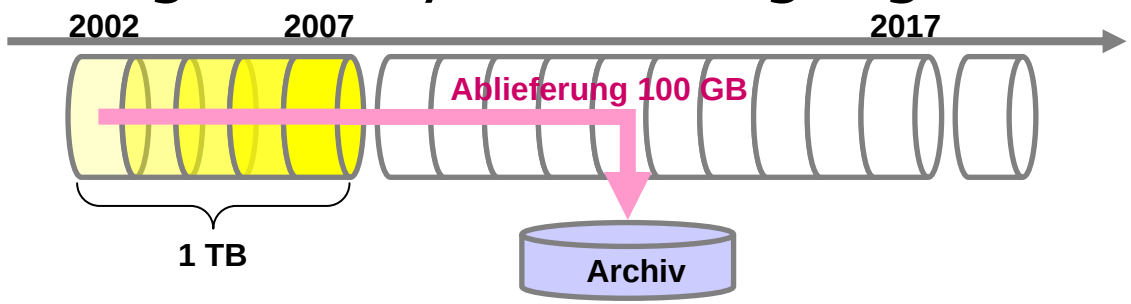
Imaging & Media Lab

- Mikrofilm, Rosetta Disk, etc. sind noch exotische Medien.

Technische Aspekte
Datenmenge

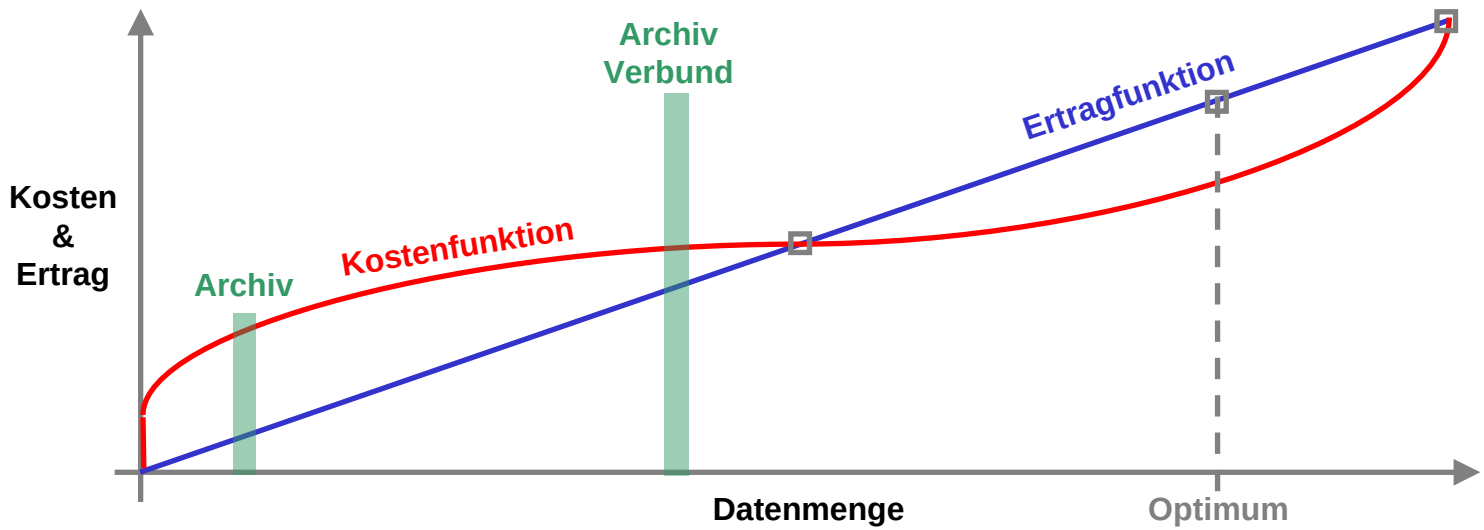
Wir müssen uns hier auf Schätzungen stützen:

- Die wenigsten kantonalen Verwaltungen haben heute mehr als 1 TB Daten online. Dieses eine TB repräsentiert wahrscheinlich etwa 5 Jahre digitale Aktenführung.
- Wenn wir nun die Hälfte der angesammelten Unterlagen als archivwürdig bewerten, ist in 10 Jahren mit 100 GB für das Jahr 2007 zu rechnen.
- Daraus folgt eine Systemauslegung von 1–10TB.



Technische Aspekte
Economy of Scale

Gerade die Datenspeicherung unterliegt einem Skaleneffekt (Ertragsgesetz der Betriebswirtschaft), deshalb sind sich alle Referenten einig, dass die Langzeitarchivierung ein Gemeinschaftsprojekt der Staatsarchive sein muss:



Organisatorische Aspekte *Zusammenarbeit*

- Für verschiedene Lösungen liegen die in einem Staatsarchiv zu erwartenden Datenmengen unter der kritischen Masse für einen wirtschaftlichen Betrieb.
- Dies ist selbstredend bei der Lösung „Speicherzentrum“ so, aber auch bei „Speicherplatz mieten“ und „Blackbox“.

Deshalb sollten Modelle der Zusammenarbeit (z.B. gemeinsames Bestellen von Leistungen) geprüft werden.

Organisatorische Aspekte
Kommerzielle Anbieter

- „Gretchenfrage“: Kommen kommerzielle Anbieter überhaupt in Frage?
  - Speicherplatz mieten
  - Speicherinfrastruktur unterbringen
- Analogie: Die klassischen Archivmagazine werden i.d.R. nicht outgesourct.
- Falls dies ein Problem sein sollte, fallen verschiedene Überlegungen von vorneherein weg.
- Zu klären sind allenfalls Definitionen und Abgrenzungen.

Organisatorische Aspekte *Rechtliches*

- Rechtliche Probleme werden als lösbar eingeschätzt; allenfalls besteht Gesetzgebungsbedarf.
- Problematisch sind allenfalls:
  - „Psychologische“ Probleme bei der Speicherung von Archivgut in einem anderen Archiv oder bei einem kommerziellen Anbieter
  - Politischer Druck, einen verwaltungsinternen Anbieter von Speicherplatz zu berücksichtigen (Informatikdienst der Verwaltung)

Kosten
Lifecycle Management

Die Speicherbewirtschaftung unterliegt in der Langzeitarchivierung speziellen Anforderungen welche auch auf die Kostenstruktur entsprechenden Einfluss hat:

- Ein Archivsystem muss nicht hochverfügbar sein.
- Zugriffszeiten ~1 min sind tolerierbar.
- Keine Mutationen oder Löschen in den Beständen.
- Der Zugriff auf die Daten erfolgt arbiträr.
- Die Integrität nicht benutzter Datenbestände muss regelmässig überprüft werden.

Kosten
Kostenfolgen

1. Keine Hochverfügbarkeit → Verzicht auf eine Reihe von Rechenzentrum Infrastrukturmassnahmen *(kein Notstromaggregat, kein Picket etc.)*
2. Lange Zugriffzeiten → *Nearline Speicherung* (z.B. nur auf Magnetband)
3. Keine Mutationen oder Löschen in den Beständen → *einfache Backupstrategie*
4. Der Zugriff auf die Daten erfolgt arbiträr → *Datenauslagerung ist nicht möglich*
5. Die Integrität muss regelmässig überprüft werden → *Storage Management Software*

Punkt 1 bis 3 senken, 4 und 5 erhöhen die Kosten!

Kosten
Digitale Speicherung-Papiermagazin

Ein Kostenvergleich zeigt folgendes:

- Beispiel: Vollkosten für ein Lager von 10'000 m2 betragen 250'000.- pro Jahr. Dieses Lager fasst etwa 10 Lkm Papierakten.
- Wir verwenden folgende Umrechnungsformel: Papierarchivgut zu digitale Daten 10 Lkm = 3 TB
  [1 Dokument à 4 Seiten ~ 50 KB => 1 Lm (6000 Seiten) = 300 GB]
- Speicherung von 3 TB nach der Vollkosten-rechnung einer kantonalen Informatikdienststelle zwischen 90'000.- und 180'000.- pro Jahr.

Digitale Speicherung ist günstiger als konventionelle, wir erwarten aber rapide steigende digitale Datenmengen!

Kosten
Kostenvergleich

Ein Kostenvergleich von *Bitstream Preservation* Anbietern zeigt grosse Unterschiede (Angaben sind schwer erhältlich und selten verbindlich):

- Verbindliches Angebot eines kantonalen Informatik-dienstleisters (30 SFr./GB/Jahr, 10 J. Laufzeit)
  30‘000 SFr./TB/Jahr
- Kauf einer Blackbox (Pauschalpreis 100‘000€/8TB/3Jahre)
  6‘850 SFr./TB/Jahr
- Kauf einer Tapelibrary (140‘000 SFr./40TB/3Jahre)
  4‘800 SFr./TB/Jahr (Hard %25 Gesamtkosten = 1‘200 SFr.)
- Günstigster Anbieter von Speicherplatz in unserem Kolloquium (0.23 SFr./GB/Monat, 5 Jahren Laufzeit)
  2‘760 SFr./TB/Jahr
- Amazon S3 (0.20 SFr./GB/Monat)
  2‘400 Sfr./TB/Jahr

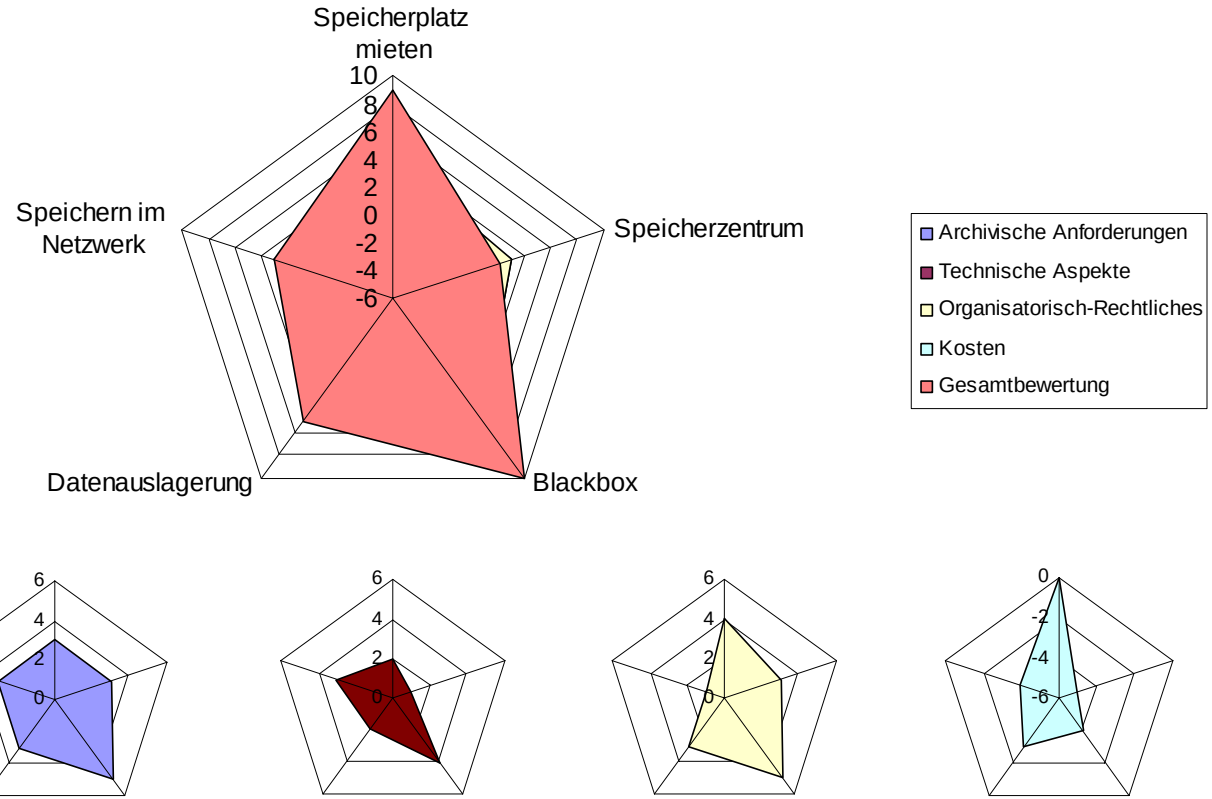
Fazit
Forschungsbedarf

- „Datenauslagerung“ und „Speichern im Netzwerk“ sind aus archivischer Sicht interessant, aber noch nicht einsetzbar:
  - fehlende Produktreife
  - fehlende Standardisierung
- Forschungsprojekte und private Anbieter sind an diesen Themen dran.
- Es könnte sich für die Archivwelt lohnen, sich hier ebenfalls zu engagieren:
  - Verstärkter Kontakt mit Forschungsinstitutionen
  - Mitarbeit bei Pilotimplementierungen
- Rolle der KOST?

Fazit
Entscheidungsmatrix I

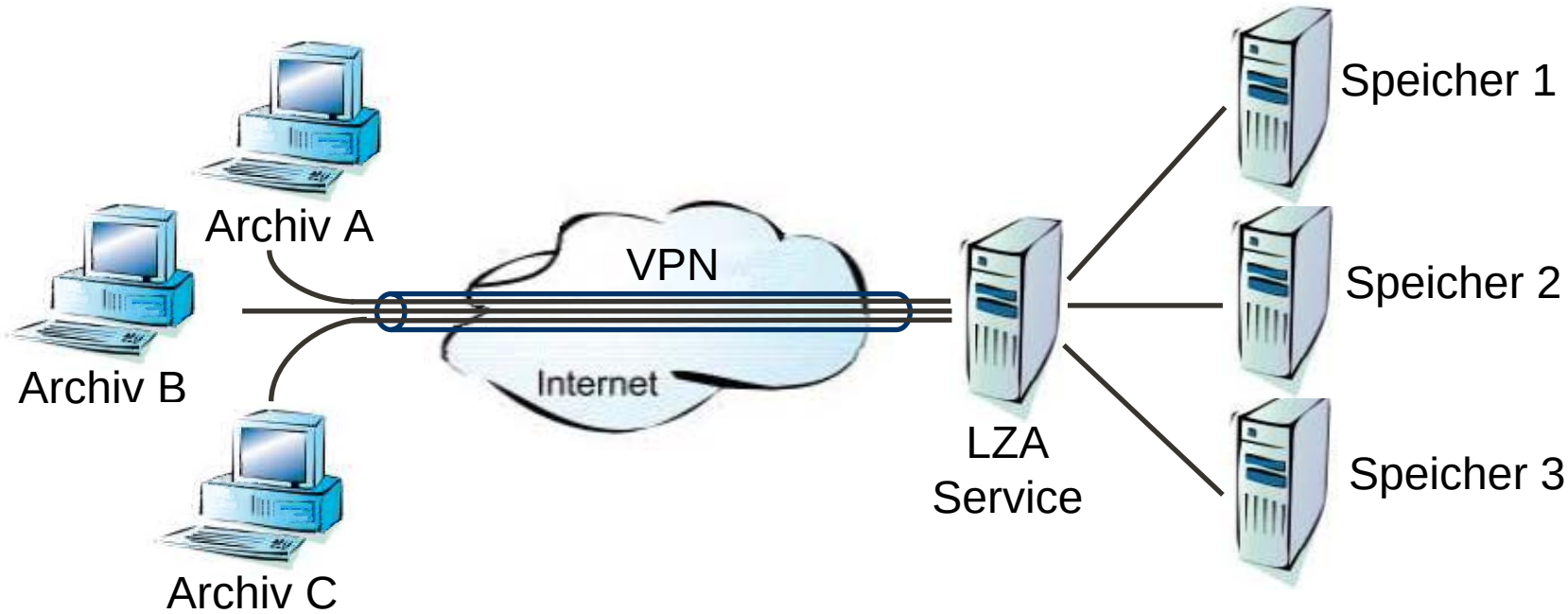
Alternativen	Speicherplatz mieten	Speicherzentrum	Blackbox	Datenauslagerung	Speichern im Netzwerk
Archivische Anforderungen					
• Archivbetrieb	2	2	3	0	1
• Langzeitarchivierung	1	1	2	3	2
Technische Aspekte					
• Innovationspotentials	0	0	1	2	3
• Best Practice	2	1	3	0	0
Organisatorisch-Rechtliches					
• Organisatorisches	3	1	2	1	1
• Rechtliches	1	2	3	2	0
Kosten					
• Investitionen	0	-3	-3	-2	-1
• Administration	0	-2	-1	-1	-3
Bewertung	**9**	**2**	**10**	**5**	**3**

Fazit
Entscheidungsmatrix II



Szenario
Speicherplatz mieten I

1. Etwa 10 Archive geben der KOST den Auftrag zur LZA von je 1TB im *Dark Archive* Modus.



Szenario
Speicherplatz mieten II

2. Die KOST spezifiziert für den LZA-Bereich eine Speicherschnittstelle (S3 oder dCache) und bestimmt eine Service Level (dreifache Redundanz, periodischer Integritätscheck, Mandantenfähigkeit).
3. Die KOST holt Offerten für den LZA-Bereich bei kommerziellen oder öffentlich-rechtlichen Anbietern (BAR, Phonothek, etc.) ein. (allenfalls WTO-Ausschreibung?)
4. Die KOST übernimmt Auditing und Inkasso für die gewählte Lösung.
5. Jedes beteiligte Archiv bezahlt pro Jahr den Minimalbetrag für 1TB (oder mehr). Kalkulations-grundlage: 3 mal 1500 SFr./TB/Jahr=4‘500 SFr.

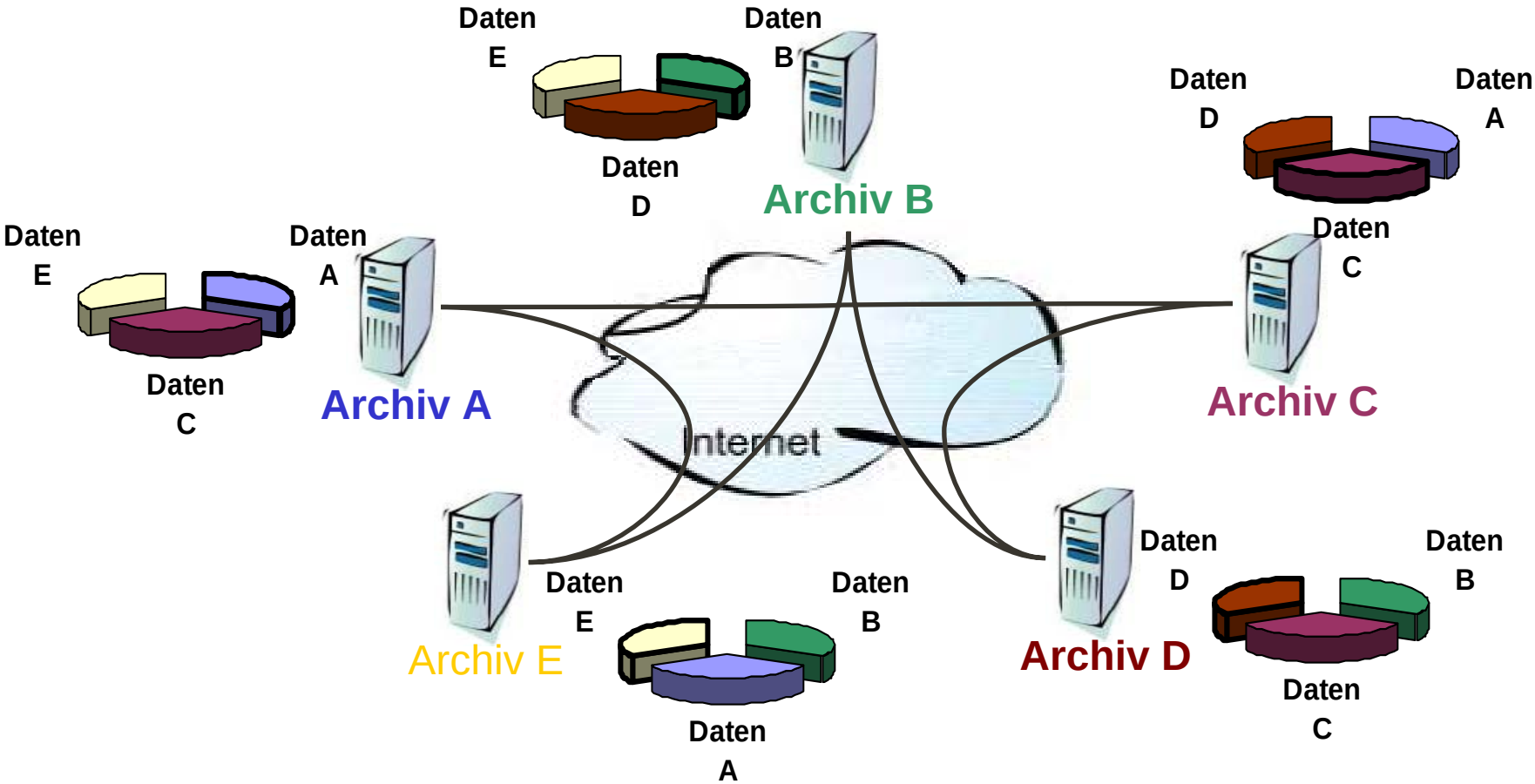
Szenario
Speicherbox teilen

1. Wie *Speicherplatz mieten,* aber es findet sich keine hinreichend interessante Offerte im Bereich Speicherplatz mieten.
2. Wie *Speicherplatz mieten.*
3. a) Die KOST kauft eine Speicherinfrastruktur vom Typus Blackbox und lässt diese in einem Rechenzentrum betreiben.
   b) Die KOST evaluiert Software für die Speicherschnittstelle (S3 oder dCache), die Replikation und den Integritätscheck.
4. Die KOST übernimmt Auditing und Inkasso.
5. Wie *Speicherplatz mieten.*

Szenario
Archivnetzwerk I

1. Die interessierten Archive kaufen eine eigene Blackbox-Lösung und betreiben diese im Archiv oder bei ihrem Informatikdienstleister.
2. Die Archive stellen $^2/_3$ ihrer Speicherkapazität anderen Archiven als *Dark Archive* zur Verfügung.
3. Die KOST evaluiert eine Softwarelösung für die Replikation der Daten von Archiv zu Archiv.
4. Die KOST organisiert die Struktur des Netzwerks und eine Lösung für die sichere Übertragung der Daten via VPN.
5. Der Datenzugriff wird vertraglich zwischen den beteiligten Archiven geregelt.

Szenario
Archivnetzwerk II



Szenario
Kalkulation I

Folgende Prämissen liegen unserem Kalkulations-beispiel zugrunde:

1. Verrechenbarer Preis für ein TB/Jahr (eine Kopie):
   1‘500 SFr.
2. Folgende Kostenverteilung wird angenommen:

Hardware/Medien	*25%*
Basisinfrastuktur	*15%*
Lizenzen & Unterhalt	*15%*
Nutzräume	*10%*
Administration	*35%*

3. Kommunikationskosten zwischen *Dark Archive* und Archiv sind nicht in der Rechnung enthalten.

Szenario
Kalkulation II

4. Hardware-Amortisation in 3 Jahren.
5. 10 Archive beteiligen sich mit 1TB:
   *10TB * 3 = 30TB*
6. Kosten pro Jahr:
   *30 * 1‘500 SFr. = 45‘000/Jahr*
7. Hardwarekosten in der Amortisationsperiode:
   *3 * 45’000 = 135‘000*
   *135‘000 * 0.25 = 33‘750*
8. Hardwarekosten aufgeteilt auf 3 Einheiten, Kosten pro Speichereinheit à 10TB:
   ~ 10‘000 SFr.