1

DEUTSCHE NATIONAL BIBLIOTHEK

# Daten, Daten, Daten und kein Ende ?

Dr. Peter Leinen

Deutsche Nationalbibliothek

2

DEUTSCHE NATIONAL BIBLIOTHEK

## Daten, Daten, Daten und kein Ende ?

oder

## Die unendlichen Weiten des Datenuniversums

DEUTSCHE NATIONAL BIBLIOTHEK

## … eintauchen die schier unendlichen Weiten

– der Datenwelt, die
  - von einer reinen Geschäftsgrundlage zu einer kritischen Größe in unserem Leben geworden ist
  - immer neue Rekorde bricht
  - immer neue Datenquellen integriert
  - immer mehr Speicherplatz erfordert



DEUTSCHE NATIONAL BIBLIOTHEK

## Gliederung

– Allgemeine Entwicklung

– Auftrag der Deutschen Nationalbibliothek

– Speichertechnologien an der Deutschen Nationalbibliothek

DEUTSCHE NATIONAL BIBLIOTHEK

# Entwicklung

- Vor 1980: Mainframe Ära
  - Datenspeicher und Rechenleistung ausschließlich in Datenzentren
  - Datensilos, Rechenleistung steht im Vordergrund
- 1980 – 2000: PC Ära
  - Daten und Rechenleistung werden dezentralisiert
- ab 2000: Cloudzeitalter
  - Daten stehen im Vordergrund, Vernetzung der Daten
  - Zugriff überall und immer
  - Grundlage: Netzwerke

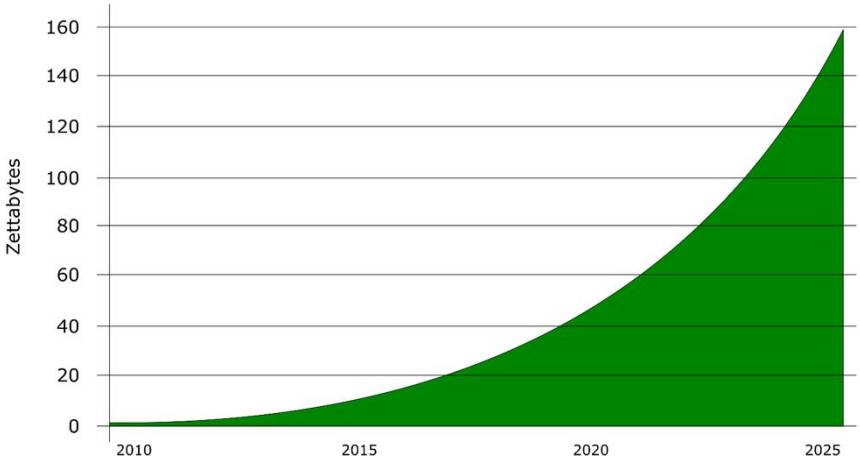


DEUTSCHE NATIONAL BIBLIOTHEK

# Datenvolumen

2018: 34 ZB

2025: 160 ZB




Quelle IDC: Data Age 2025, April 2017

DEUTSCHE NATIONAL BIBLIOTHEK

## Größenordnungen

| Einheit | | | Faktor |
|---|---|---|---|
| 1 KB | Kilobyte | 1.000 Bytes | 1000 |
| 1 MB | Megabyte | 1.000.000 Bytes | $1000^2$ |
| 1 TB | Terabyte | 1.000.000.000 Bytes | $1000^3$ |
| 1 PB | Petabyte | 1.000.000.000.000 Bytes | $1000^4$ |
| 1 EB | Exabyte | ... | $1000^5$ |
| 1 ZB | Zettabyte | ... | $1000^6$ |
| 1 YB | Yottabyte | ... | $1000^7$ |

Unterschied: 1 KiB = 1024 B, 1 TiB = 1024 MiB = $1024^2$ B



DEUTSCHE NATIONAL BIBLIOTHEK

## Fragen

- Wo entstehen diese Datenmengen?
- Wo werden diese gespeichert?
- Welche Technologien werden hierzu benutzt?

DEUTSCHE NATIONAL BIBLIOTHEK

## Beispiel: Large Hadron Collider (LHC)

- Teilchenbeschleuniger am Kernforschungszentrum CERN
  - Erzeugung und genaue Untersuchung von Elementarteilchen
- Datenverarbeitung
  - mehr als 99% Messungen (Standardsignaturen) werden gleich verworfen
  - „nur“ 200 Messwerte pro Sekunde werden weitergeleitet
  - Datenproduktion: 25 PB / Jahr



DEUTSCHE NATIONAL BIBLIOTHEK

## Beispiel: Large Hadron Collider (LHC)

- LHC Computing Grid
  - Netz von Einrichtungen zur Speicherung und Prozessierung
- Tier-0: CERN
  - Backup auf Tape
  - Verarbeitung mit 6000 CPUs, 5.5 PB Festplatte, 17 PB Tape
- Tier-1: 13 weltweit
  - Großforschungseinrichtungen (KIT Karlsruhe)
  - jeweils 2000 CPUs, 1 PB Festplatte, 10 PB Tape

DEUTSCHE NATIONAL BIBLIOTHEK

## Beispiel: Large Hadron Collider (LHC)

- Tier-2: 155 weltweit
  - Universitäten, Forschungseinrichtungen
  - spezielle Auswertungen
- Tier-3, Tier-4:
  - Forschergruppen



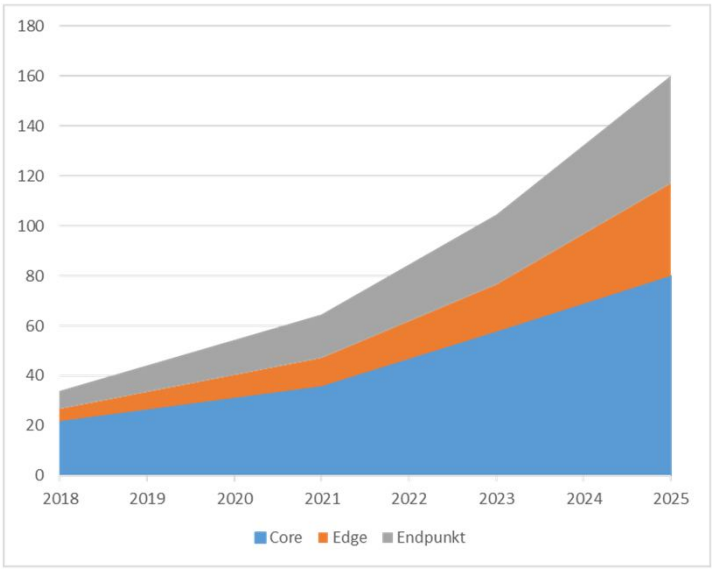
DEUTSCHE NATIONAL BIBLIOTHEK

## Datenquellen

- Zentral (Core)
  - Große Datenzentren
- Endpunkte
  - PCs, Telefone, Kameras
  - Fahrzeuge, Sensoren
- Edge
  - Außenstellen, Netzwerkserver
  - Verbindung zwischen Zentral und Endpunkten

DEUTSCHE NATIONAL BIBLIOTHEK

## Datenquellen






DEUTSCHE NATIONAL BIBLIOTHEK

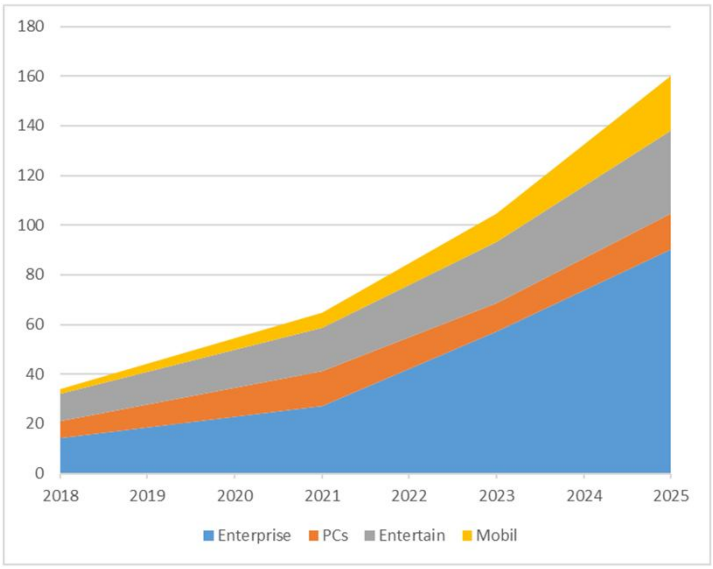
## Datenquellen

| % | 2018 | | 2021 | 2023 | 2025 |
|---|---|---|---|---|---|
| Core | 64 | | 55 | 55 | 50 |
| Edge | 14 | | 18 | 18 | 23 |
| Endpunkt | 21 | | 27 | 27 | 27 |

| PB | 2018 | | 2021 | 2023 | 2025 |
|---|---|---|---|---|---|
| Core | 22 | | 36 | 58 | 80 |
| Edge | 5 | | 12 | 19 | 37 |
| Endpunkt | 7 | | 17 | 28 | 43 |

DEUTSCHE NATIONAL BIBLIOTHEK

## Speicherorte






DEUTSCHE NATIONAL BIBLIOTHEK

## Speicherorte

| % | 2018 | | 2021 | 2023 | 2025 |
|---|---|---|---|---|---|
| Enterprise | 42 | | 42 | 55 | 56 |
| PCs | 20 | | 22 | 11 | 9 |
| Entertain | 33 | | 27 | 24 | 21 |
| Mobil | 5 | | 9 | 11 | 14 |

| PB | 2018 | | 2021 | 2023 | 2025 |
|---|---|---|---|---|---|
| Enterprise | 14 | | 27 | 57 | 90 |
| PCs | 7 | | 14 | 11 | 15 |
| Entertain | 11 | | 18 | 25 | 33 |
| Mobil | 2 | | 6 | 11 | 22 |

DEUTSCHE NATIONAL BIBLIOTHEK

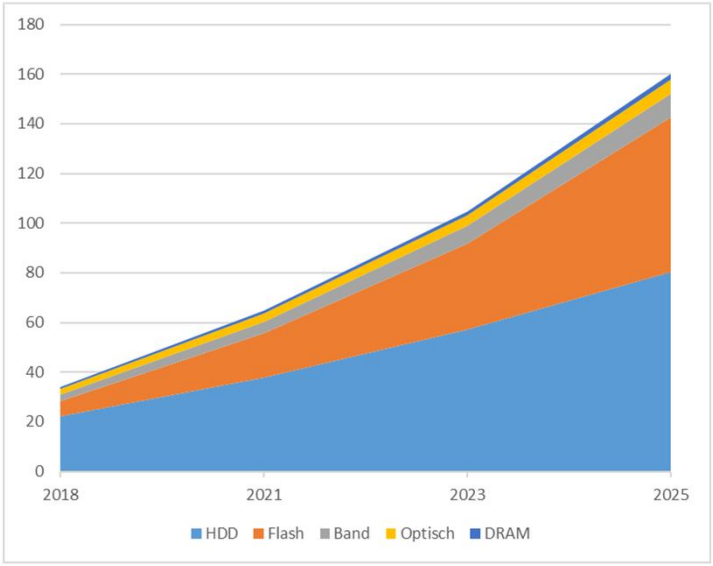
## Technologien

| % | 2018 | | 2021 | 2023 | 2025 |
|---|---|---|---|---|---|
| HDD | 66 | | 59 | 55 | 50 |
| Flash | 17 | | 27 | 33 | 39 |
| Band | 8 | | 7 | 7 | 6 |
| Optisch | 7 | | 5 | 4 | 4 |
| DRAM | 1 | | 1 | 1 | 1 |



DEUTSCHE NATIONAL BIBLIOTHEK

## Technologien



DEUTSCHE NATIONAL BIBLIOTHEK

## Technologien

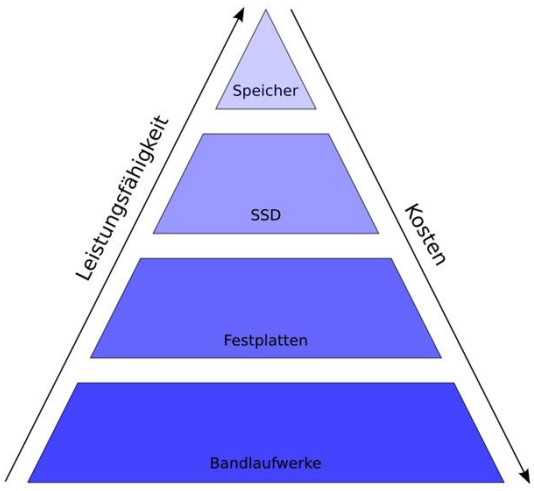
| PB | 2018 | | 2021 | 2023 | 2025 |
|---|---|---|---|---|---|
| HDD | 22 | | 38 | 57 | 80 |
| Flash | 6 | | 18 | 34 | 62 |
| Band | 3 | | 5 | 7 | 10 |
| Optisch | 2 | | 4 | 4 | 6 |
| DRAM | 0,4 | | 1 | 1 | 2 |



DEUTSCHE NATIONAL BIBLIOTHEK

## Speicherhierarchie

- Festplatten (HDD/SSD)
  - hohe Performance
  - hohe Betriebskosten
- Bandlaufwerke
  - selten genutzt
  - wenig Veränderung
  - Langzeitarchivierung
- Gemeinsamkeit
  - Endliche Haltbarkeit



DEUTSCHE NATIONAL BIBLIOTHEK

## Quelle

Data Age 2025: The Evolution of Data to Life-Critical

https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf



DEUTSCHE NATIONAL BIBLIOTHEK

## Deutsche Nationalbibliothek

- Aufgaben
- Digitaler Wandel
- Anforderungen

DEUTSCHE NATIONAL BIBLIOTHEK

## Deutsche Nationalbibliothek

– Archivbibliothek mit der Aufgabe,
  - die in Deutschland ab 1913 veröffentlichten Medienwerke,
  - im Ausland veröffentlichten deutschsprachigen Medienwerke, Übersetzungen deutschsprachiger Medienwerke in andere Sprachen und fremdsprachige Medienwerke über Deutschland
  - die zwischen 1933 und 1945 erschienenen Werke deutschsprachiger Emigranten

  zu sammeln, dauerhaft zu archivieren, bibliografisch zu verzeichnen sowie der Öffentlichkeit zur Verfügung zu stellen.



DEUTSCHE NATIONAL BIBLIOTHEK

## Deutsche Nationalbibliothek

– **2006:** DNB-Gesetz
  - Umbenennung in Deutsche Nationalbibliothek
  - **Ausweitung Sammelauftrag auf Netzpublikationen**

– Netzpublikationen (Online-Publikationen)
  - E-Books, elektronische Zeitschriften
  - Hörbücher
  - Webseiten

DEUTSCHE NATIONAL BIBLIOTHEK

## Bestand am 1.1.2019

- Ca. 30 Mio. physische Medien:
  - Bücher, Zeitschriften, Dissertationen, Karten, Musikalien, Tonträger
- Ca. 5.9 Mio. digitale Medien (NP):
  - E-Books, E-Journals, E-Papers, Books on Demand (BoD),
  - digitale Musiknoten, Audio-Books als Audiofiles,
  - Webseiten
- Platz
  - 380 Regalkilometer
  - 90 TB



DEUTSCHE NATIONAL BIBLIOTHEK

## Bestand am 1.1.2019

- Netzpublikationen im Einzelnen

| | | |
|---|---|---|
| 960.000 | E-Books | |
| 620.000 | BoD-Titel | |
| 230.000 | Online-Dissertationen | |
| 1.900.000 | E-Journal-Ausgaben | (7.500 Titel, 7.300 laufend) |
| 2.100.000 | E-Paper-Ausgaben | (1.450 Titel, 1.300 laufend) |
| 6.200 | Digitale Musiknoten | |
| 26.000 | Audio-Books als Audiofiles | |
| 17.400 | “Zeitschnitte” von 3047 Websites | |

- Gesamt: 5.900.000

DEUTSCHE NATIONAL BIBLIOTHEK

## Zugang

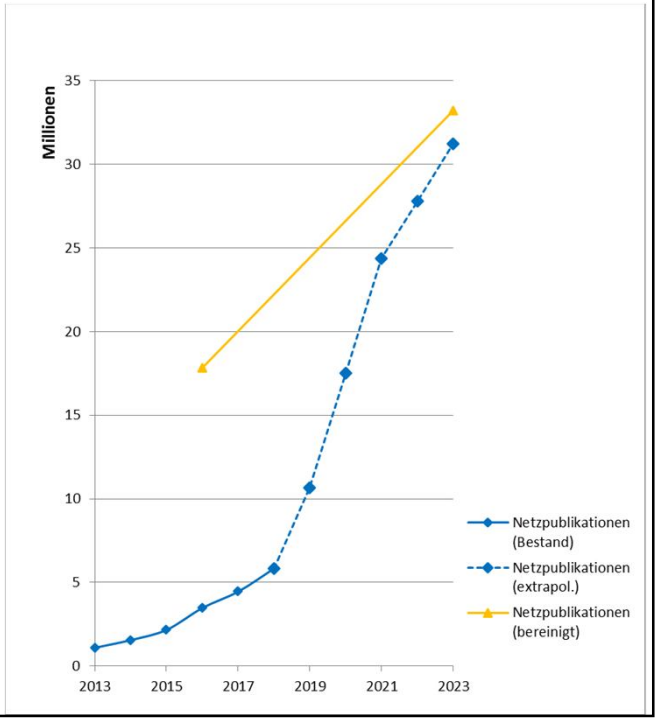
- 2017
  - Digital: 982.102
- 2018
  - Digital: 1.399.999

- Bis zu 6.000 Objekte pro Tag über diverse Schnittstellen



## IT-Strategie

- Netzpublikationen
  - Rückstand 15 Mio. NP
  - Neu 2 bis 3 Mio./Jahr
- Perspektive 2023:
  - 32 Mio. NP
  - ~ 1 PB Speicher
- Herausforderungen
  - Webarchivierung
  - Digitale Musik



DEUTSCHE NATIONAL BIBLIOTHEK

# Webarchivierung

- Archivierung „Deutsches Internet“
- „.de-crawl“ in 2014
  - 6 Mio. Websites
  - 2,5 Mrd. Dateien
  - 120 TB
- 16 Mio. Websites mit Endung „.de“
- Repräsentativität
  - 250 TB/Jahr



DEUTSCHE NATIONAL BIBLIOTHEK

# Verarbeitung NP

- Architektur
  - Parallelisierung
  - Automatisierung
- Prozesse
  - NP-Import
  - Webarchiv
  - LZA
  - Bereitstellung

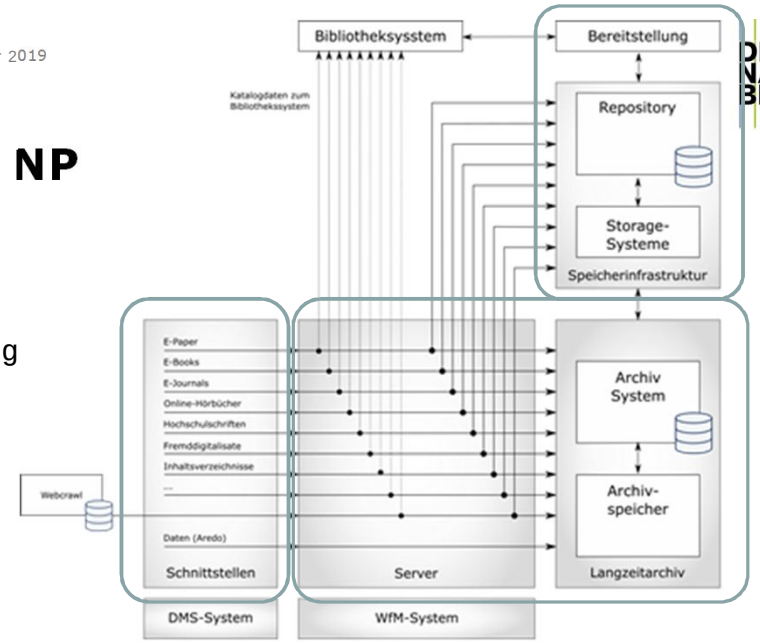
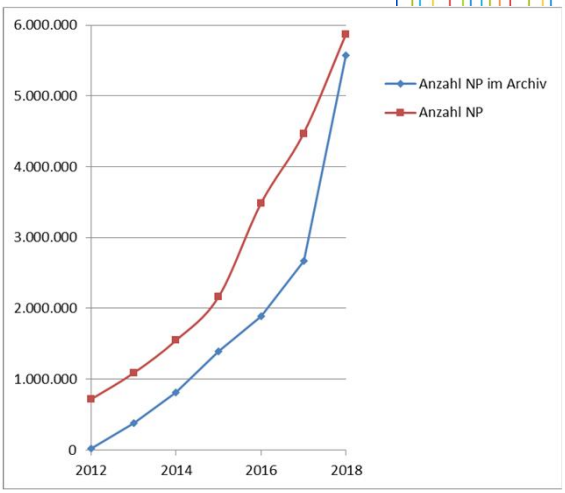



Abbildung 3: Geplante Gesamtarchitektur Netzpublikationen und LZA

DEUTSCHE NATIONAL BIBLIOTHEK

# Langzeitarchivierung

- Stand
  - Zertifiziert
    - Nestor
    - Data Seal of Approval
  - Kooperation mit GWDG
  - Neue Systemsoftware 10.2017
- Zahlen ohne
  - Digitalisate
  - Webarchiv




GWDG: Gesellschaft für Wissenschaftliche Datenverarbeitung Göttingen



DEUTSCHE NATIONAL BIBLIOTHEK

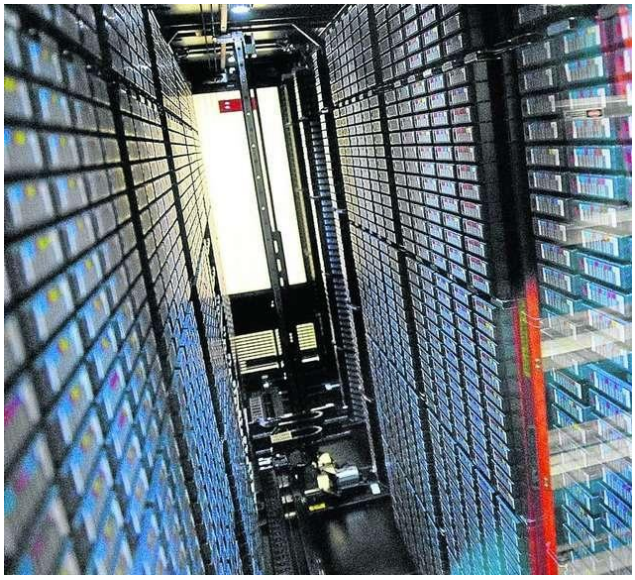
# Digitale Objekte

- Bereitstellung
  - Lesesaal
  - Text- und Datamining
  - sekundärer Speicher
  - 2-fach redundant
- Langzeitarchivierung
  - primärer Speicher
  - 3-fach redundant
  - dunkles Archiv
  - Erhaltungsprozesse

DEUTSCHE NATIONAL BIBLIOTHEK

## Bandlaufwerke

- LTO-8
  - unkomprimiert:
    - 12,8 TB/Band
    - 472 MB/s
  - komprimiert:
    - 32 TB/Band
    - 1180 MB/s
- Bedarf DNB
  - 90 – 240 Bänder



Bildquelle: GWDG



DEUTSCHE NATIONAL BIBLIOTHEK

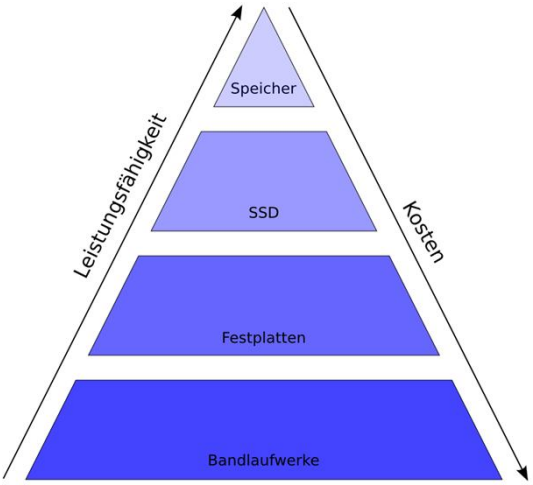
## Bandlaufwerke

- LTO-10
  - unkomprimiert:
    - 48 TB/Band
  - komprimiert:
    - 120 TB/Band
- Verdopplung der Kapazität
  - alle 2 bis 3 Jahre

DEUTSCHE NATIONAL BIBLIOTHEK

## Speichertechnologien

- Bereitstellung
  - Festplattenbasiert
  - lokaler Cache des LZA
- Langzeitarchiv
  - Bandlaufwerke
  - GWDG in Göttingen






DEUTSCHE NATIONAL BIBLIOTHEK

## Strategie der DNB

- Bereitstellung
  - häufig benötigte Objekte im direkten Zugriff
  - Speicherung an beiden Standorten
  - Speicherung auf Festplatten (HDD/SSD)
- Langzeitarchivierung
  - Speicherung durch Dienstleister
  - Bandlaufwerke aus Kostengründen
  - örtliche Redundanz