DIGITAL PRESERVATION AT THE NATIONAL LIBRARY OF NORWAY

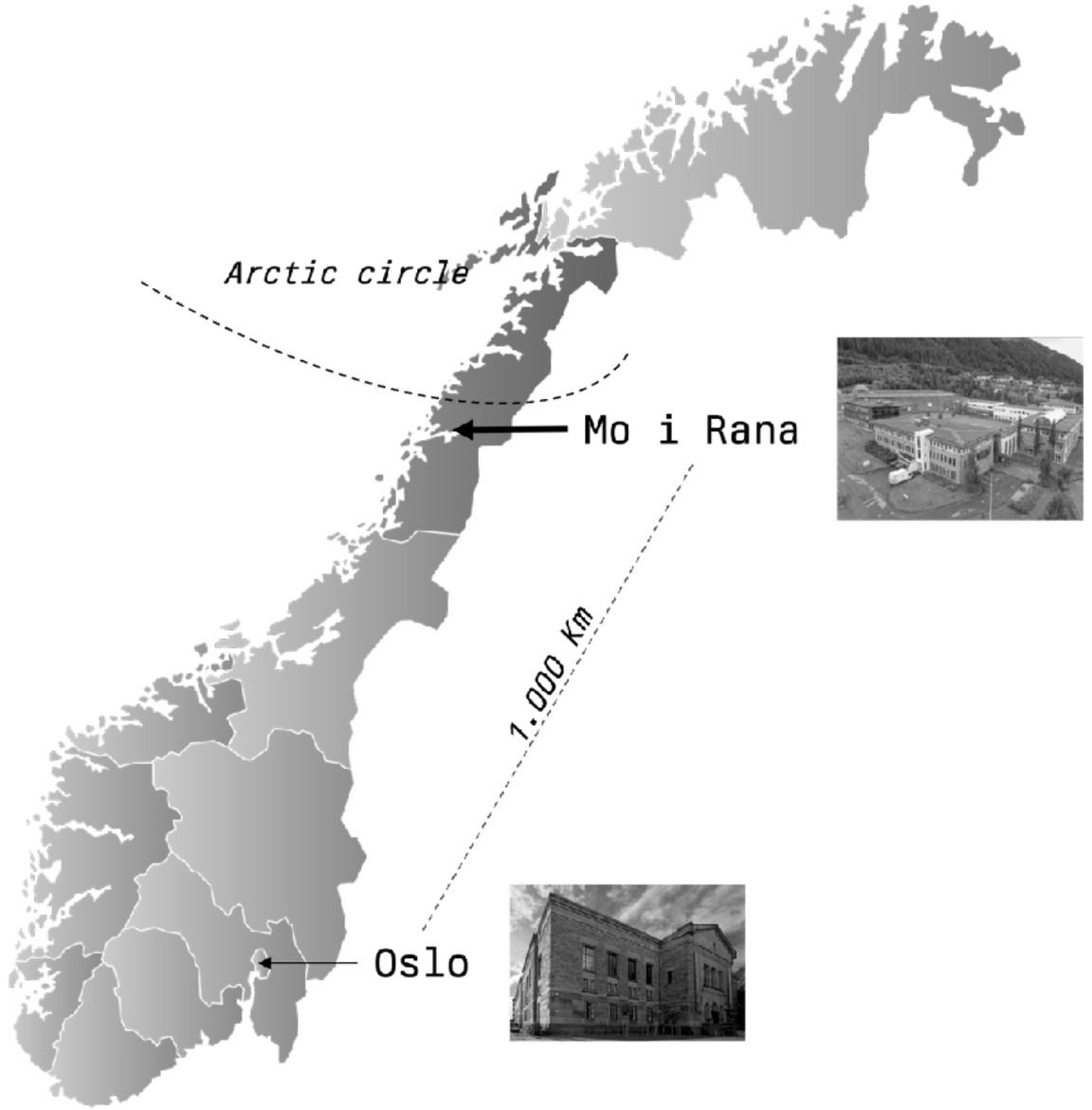
Torbjørn Pedersen 2024-29-05

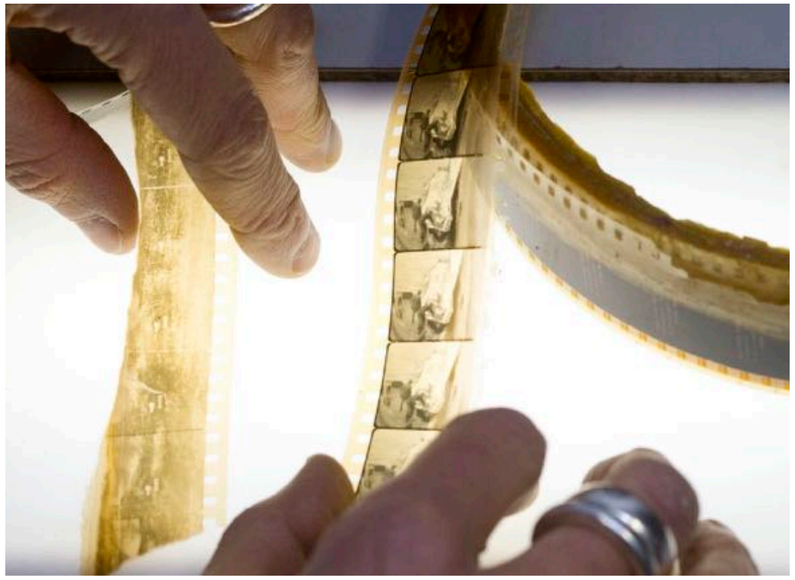
Past→present→future

From digital storage
→
to digital preservation

THE NATIONAL LIBRARY OF NORWAY (NLN)

- Almost 600 employees at 2 locations: Oslo and Mo i Rana
- **Governed by:**
  - The Legal Deposit act
- **Responsible for:**
  - Collecting, preserving, making available all content published in Norway
  - (also historical collections across all media types)






FROST
MAYHEM
AGFA
AMPEX
456

DIGITISATION PRE-2006

Early beginnings/Ancient history

- Large scale digitisation (at the time)
  - Photography
  - Audio and radio
- Formed basis for what was to come… (2006⟶)

- Basic bit storage and backups
  - **Two data loss incidents**

MASS DIGITIZATION 2006→2022

- **New mantra:**
  - “If it’s not online, it doesn’t exist”
  - “Analog objects wear out when accessed, digital objects don’t”
- **Decision:** digitize the entirety of our analog collection for preservation (and access)

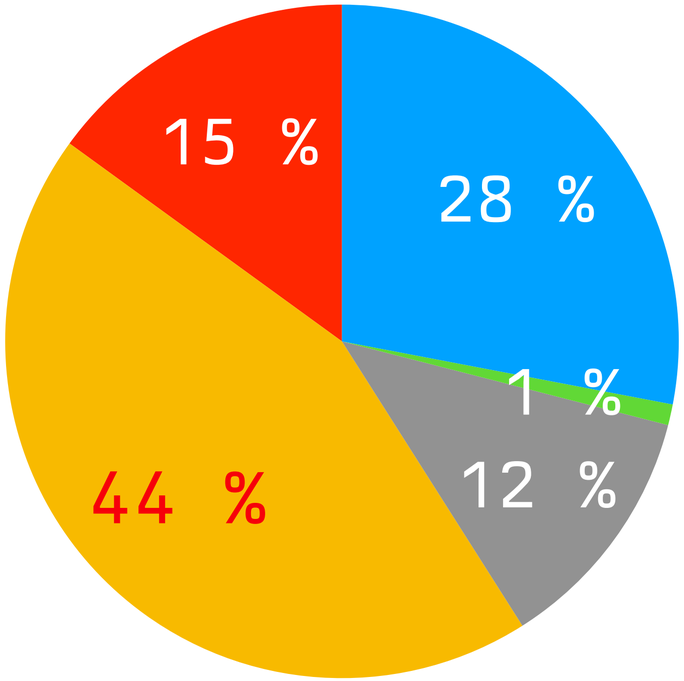
- **Mass digitisation** efforts for all media types
  - “Production lines” established - investment in people, equipment, automation
  - 30+ different production lines created over time

- Born digital deposits gradually increasing

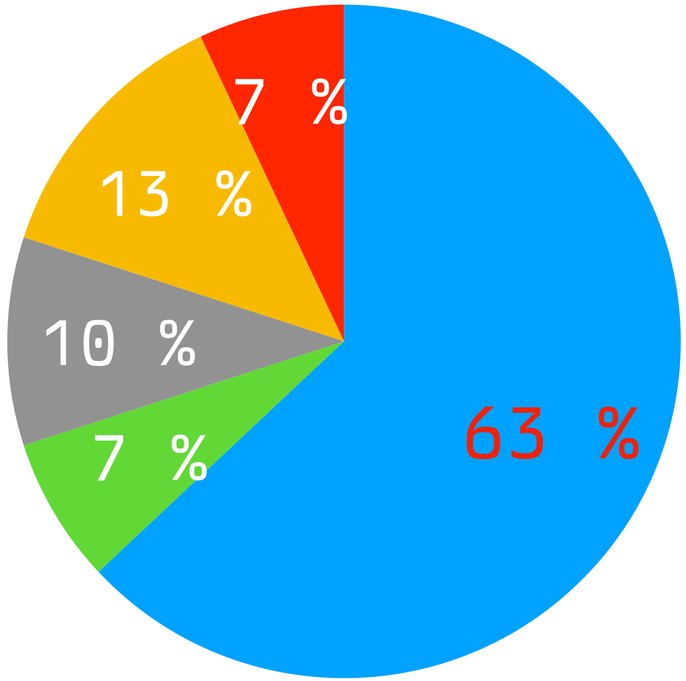
DIGITAL PRESERVATION 2006→2022

- **3-2-1 storage policy (bit preservation)**
  - 3 copies (disk+tape+tape)
  - 2 technologies: disk (luxurious!) and tape
  - 1 copy at a “off-site” location
- **New bit repository (2007):** SAM-FS (Oracle HSM)
  - Solaris OS
  - File based storage
  - Mix of different disk and tape technologies (no vendor lockin)
    - **Disk:** SUN/Oracle, Nexsan, Fujitsu, Huawei)
    - **Tape:** T10kA, T10kB, T10kC, T10kD, LTO8
  - **No (known) data loss since 2007!**

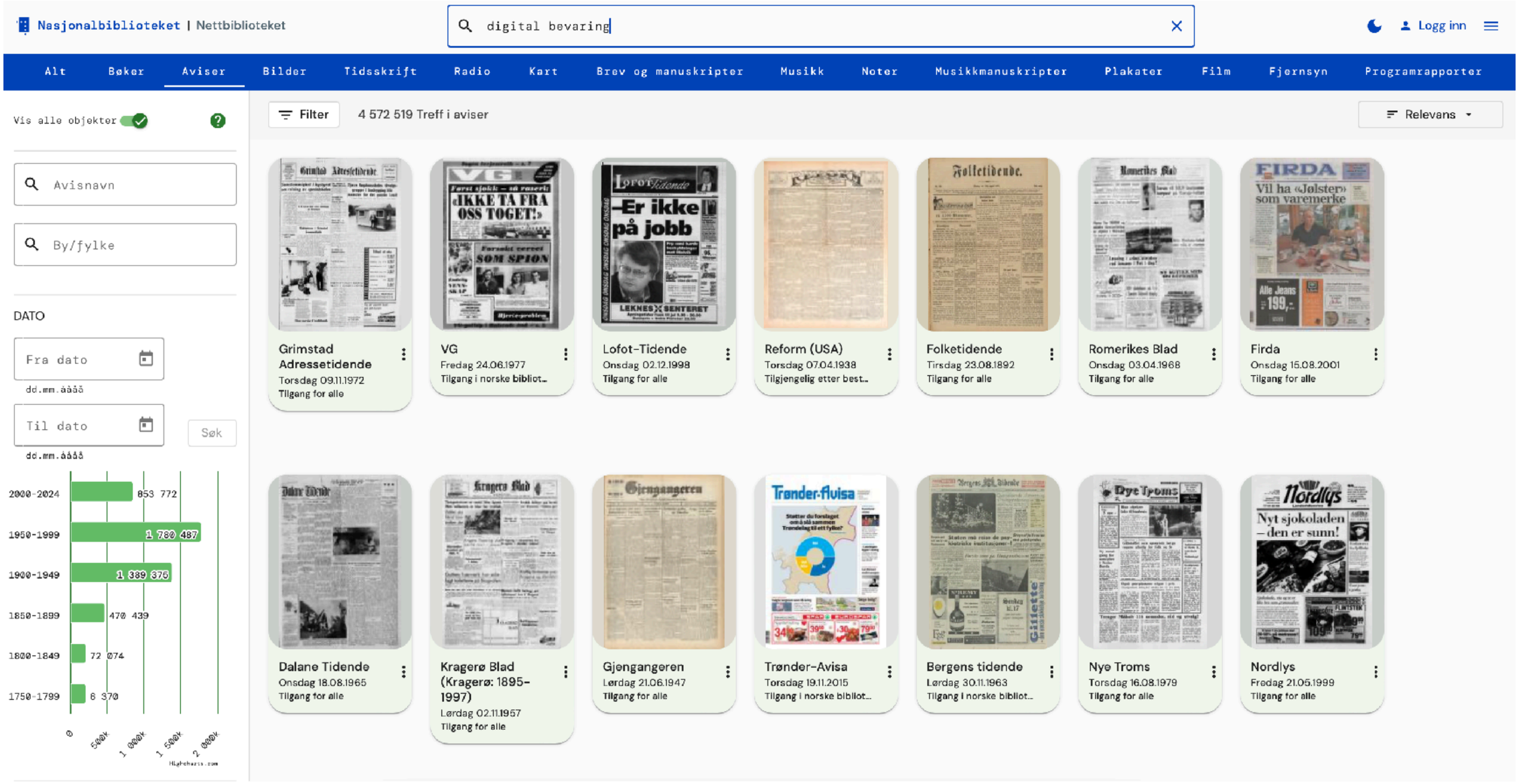
Data volume %
(16 PB)



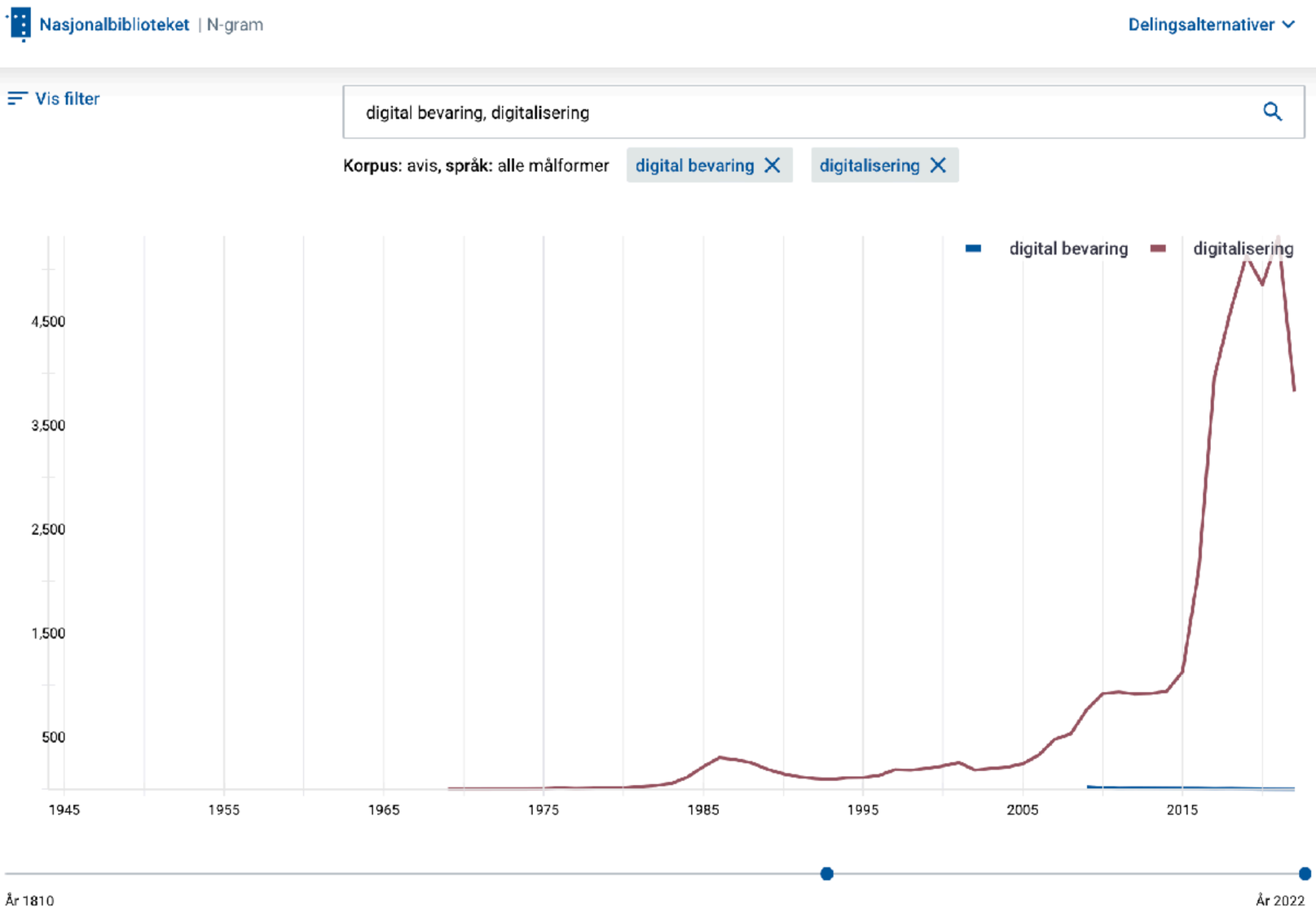
Number of files %
(2+ billion)



Media type	Data volume %	No of files %
Text	28 %	63 %
Images	1 %	7 %
Sound	12 %	10 %
Moving Images	44 %	13 %
Web harvesting	15 %	7 %


Nasjonalbiblioteket | Nettbiblioteket
digital bevaring
Logg inn
Alt
Bøker
Aviser
Bilder
Tidsskrift
Radio
Kart
Brev og manuskripter
Musikk
Noter
Musikkmanuskripter
Plakater
Film
Fjernsyn
Programrapporter
Vis alle objekter
Avisnavn
By/fylke
DATO
Fra dato
dd.mm.åååå
Til dato
dd.mm.åååå
Søk
2000-2024 853 772
1950-1999 1 780 487
1900-1949 1 389 376
1850-1899 470 439
1800-1849 72 074
1750-1799 8 370
0
500k
1 000k
1 500k
2 000k
Filter
4 572 519 Treff i aviser
Relevans
Grimstad Adressetidende
Torsdag 09.11.1972
Tilgang for alle
VG
Fredag 24.06.1977
Tilgang i norske bibliot...
Lofot-Tidende
Onsdag 02.12.1998
Tilgang for alle
Reform (USA)
Torsdag 07.04.1938
Tilgjengelig etter best...
Folketidende
Tirsdag 23.08.1892
Tilgang for alle
Romerikes Blad
Onsdag 03.04.1968
Tilgang for alle
Firda
Onsdag 15.08.2001
Tilgang for alle
Dalane Tidende
Onsdag 18.08.1965
Tilgang for alle
Kragerø Blad (Kragerø: 1895-1997)
Lørdag 02.11.1957
Tilgang for alle
Gjengangeren
Lørdag 21.06.1947
Tilgang for alle
Trønder-Avisa
Torsdag 19.11.2015
Tilgang i norske bibliot...
Bergens tidende
Lørdag 30.11.1963
Tilgang i norske bibliot...
Nye Troms
Torsdag 16.08.1979
Tilgang for alle
Nordlys
Fredag 21.05.1999
Tilgang for alle

https://nb.no

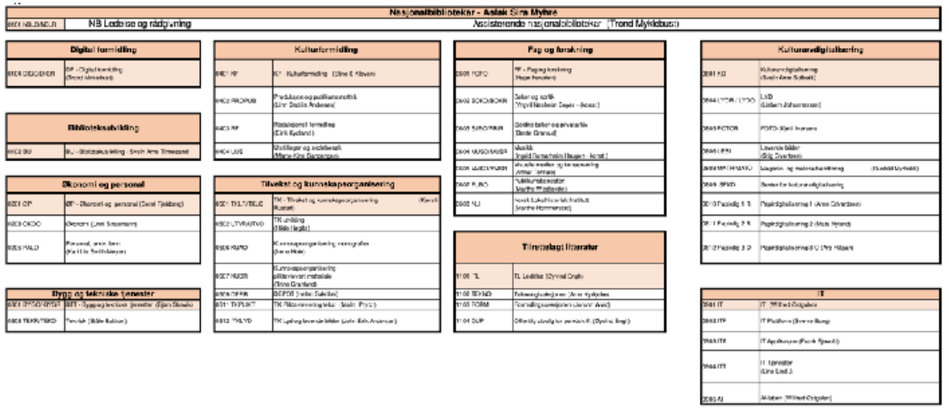


https://www.nb.no/ngram/

CHALLENGES: IT

Status →2022

- **Traditional hierarchical organization structure**
  - Decisions floating upwards
- **Myriad stakeholders**
  - Complex prioritizing/competing goals
- **IT bottleneck**
  - Solving problems in an ad hoc manner
  - Constant context switching
    - Lack of continuity



CHALLENGES: DIGITAL PRESERVATION

Status →2022

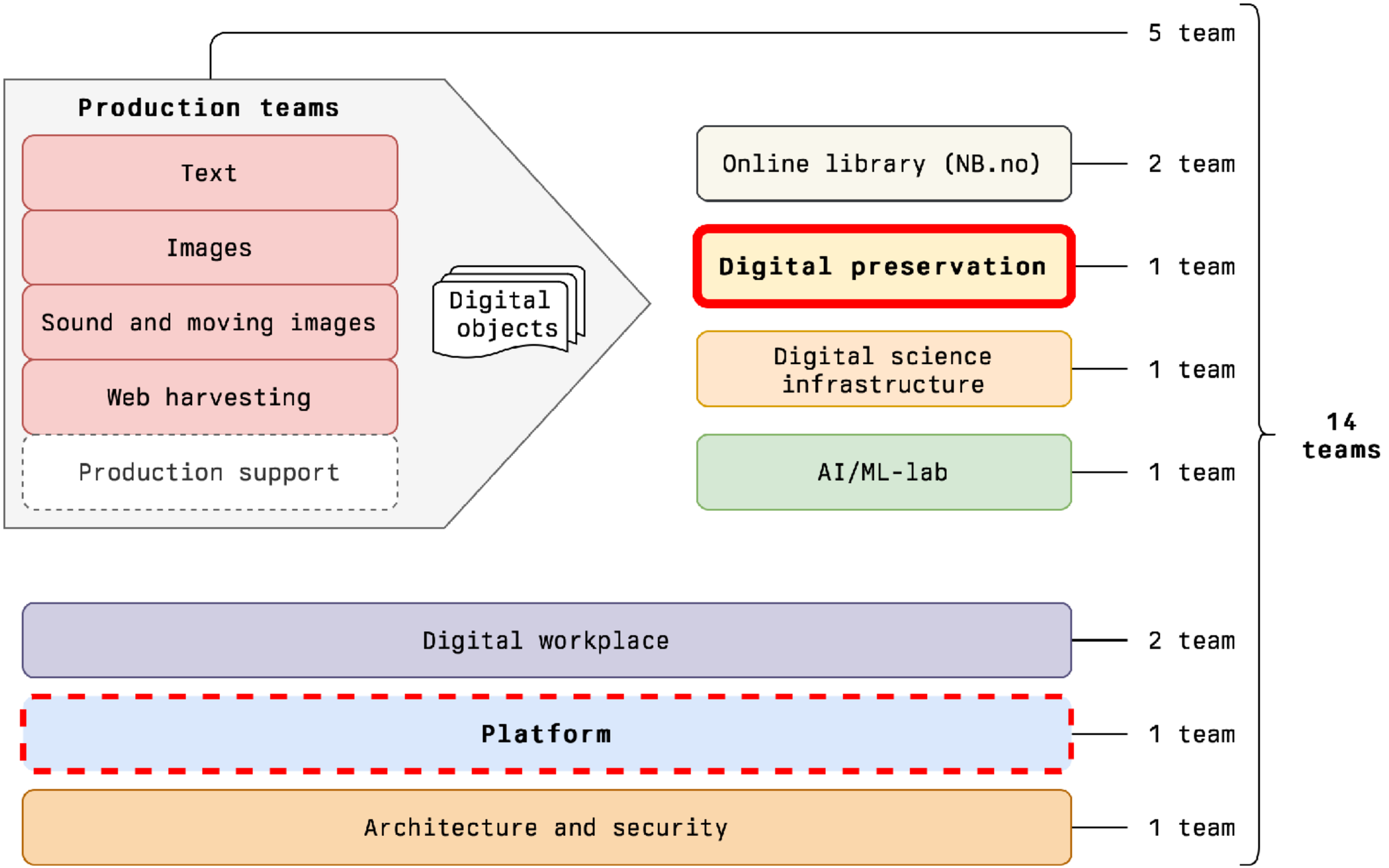
- Viewed as an IT problem - **previous issues apply**
- **Store-and-forget** mentality towards files (stored ≠ preserved!)
- Shared responsibility = **no ones responsibility!**
- **Lack of consistency:**
  - Significant amount of files lack of checksums
  - Lack of metadata
  - Data packages not standardized
  - No overall knowledge about what was stored
  - Limited knowledge about how the data was used

2022 - BIG CHANGE IN IT-STRUCTURE

New IT-director

- **Product orientation** introduced
  - Build **dynamic products** (MVP) instead of **static solutions**

- **Autonomous teams** in charge of “products”
  - **Interdisciplinary** team members (organization looks the same)
  - Product owner groups **set direction** (directors/section heads)
  - The team set their own day-to-day **priorities**

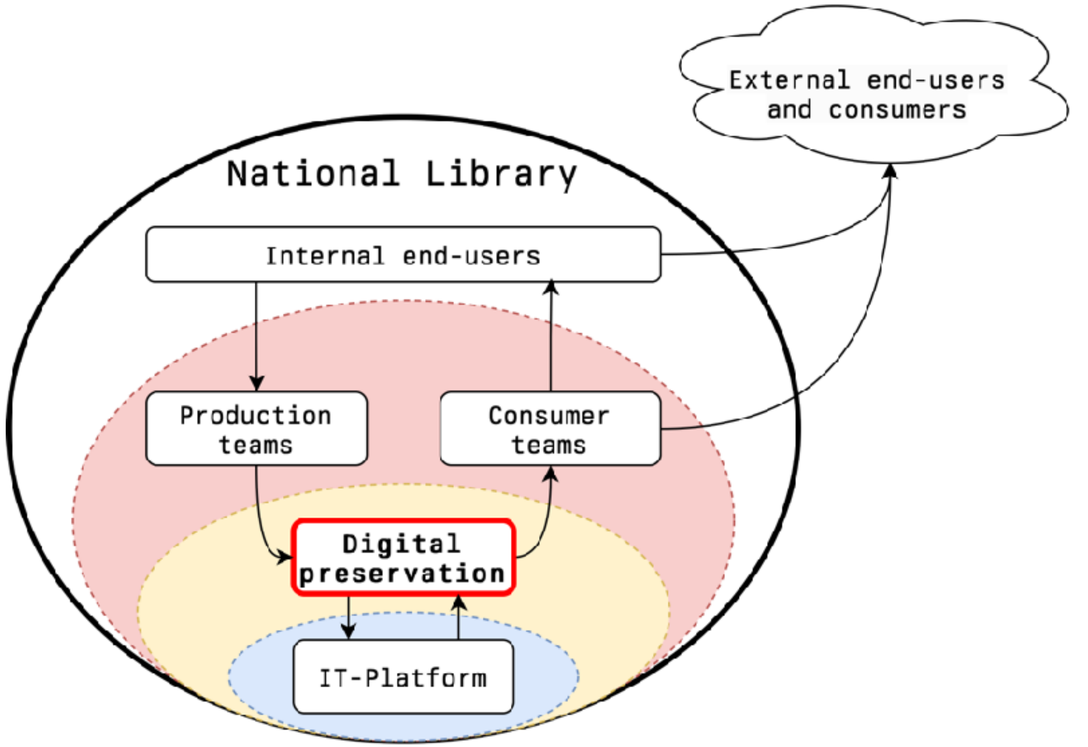

Production teams
Text
Images
Sound and moving images
Web harvesting
Production support
Digital objects
5 team
Online library (NB.no)
2 team
Digital preservation
1 team
Digital science infrastructure
1 team
AI/ML-lab
1 team
14 teams
Digital workplace
2 team
Platform
1 team
Architecture and security
1 team

DIGITAL PRESERVATION 2022→

Team established June 2022

- **Autonomous**, but not interdisciplinary (until late 2023)
- **Defined scope of responsibility**
  - Where preservation starts (the team does not create data/digitize materials)
  - Assure files+content lasts (forever)
  - Primarily an **internal** service
- **Answers to owner board**
  - Clear priorities

= Enabling systematic work with digital preservation!



Organization layers/Data flow

DIGITAL PRESERVATION 2022→

As product

Digital preservation team products:

1. **Domain expertise** of digital preservation in the organization
    - Responsibility for **building competence** and **spreading awareness**
2. **“Digital Preservation Services”** (DPS)
    - Responsibility for **developing** and **operating** the DPS-software

NLN DIGITAL PRESERVATION STRATEGY

https://digitalpreservation-blog.nb.no/docs/strategy

Ambition

- Ensure the protection of, and meaningful access to, national digital cultural heritage for current and future generations.

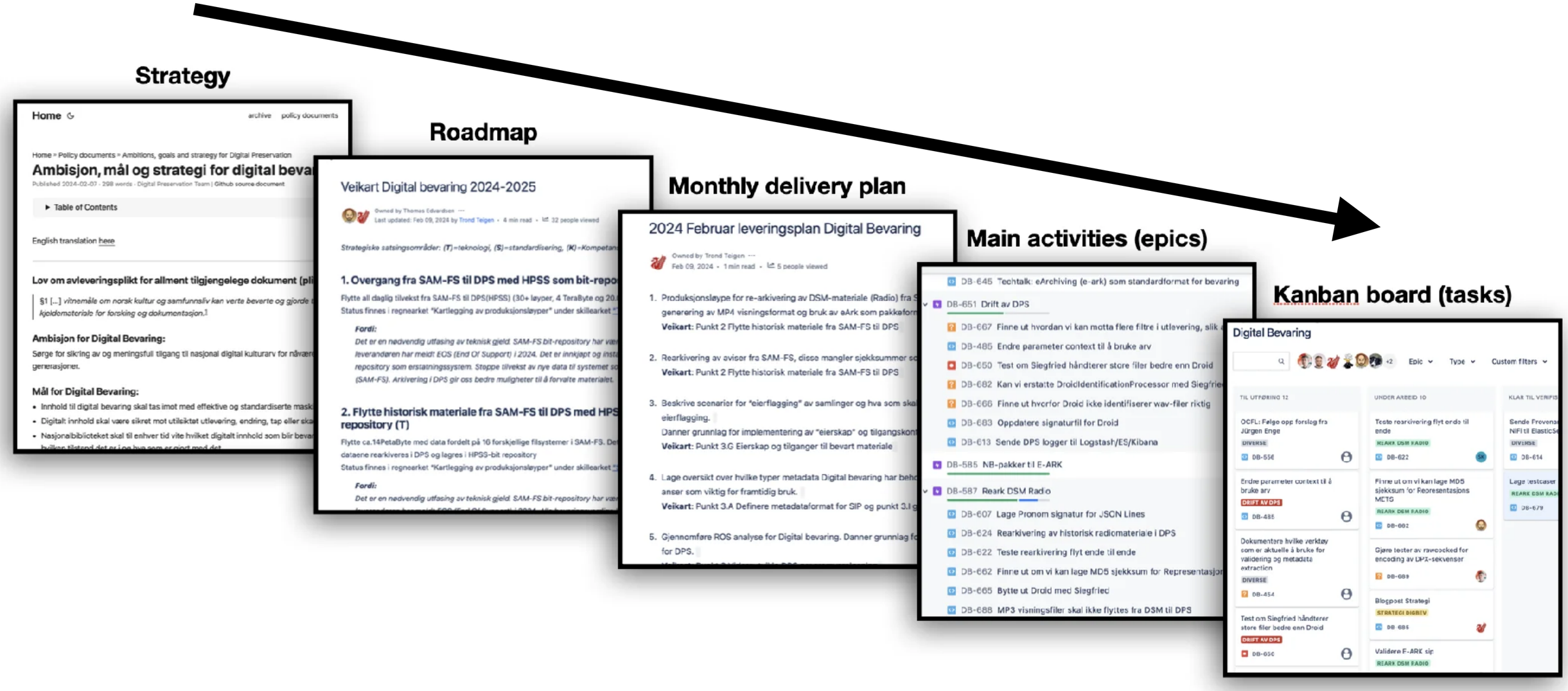
Goals for Digital Preservation

- Digital content for digital preservation shall be received using efficient and standardized machine solutions.
- Digital content shall be protected against unintended access, alteration, loss, or damage.
- The National Library shall at all times know what digital content is being preserved, its provenance, its condition, and what has been done to it.
- Digitally preserved content shall be accessible for dissemination now and in the future.

NLN DIGITAL PRESERVATION PRINCIPLES

https://digitalpreservation-blog.nb.no/docs/principles

- Ensure that digital preservation is done in a sustainable way
- Use well-documented and open file formats wherever possible
- Preserve the original file
- Analyze files that is to be preserved
- Maintain sufficient metadata to ensure that the files are identifiable and retrievable
- Use a standardized format to package files for preservation
- Standardize documentation preservation activities
- Files should be readable and understandable in the present
- Ensure that a file is stored in multiple instances, on different storage technologies and in different geographical locations (3-2-1)

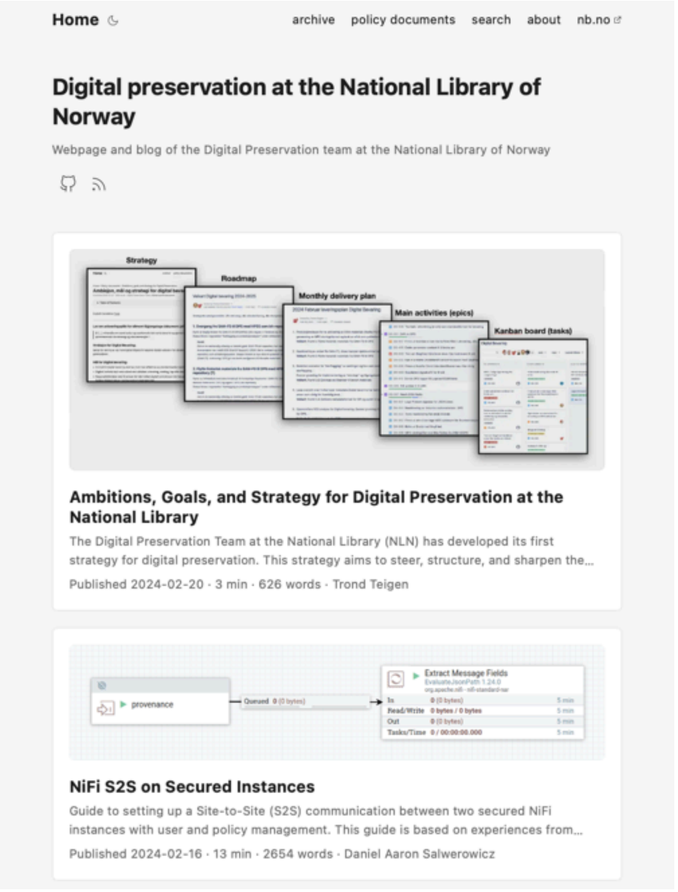

Strategy
Home
archive
policy documents
Home » Policy documents » Ambitions, goals and strategy for Digital Preservation
Ambisjon, mål og strategi for digital beva
Table of Contents
English translation here
Lov om avleveringsplikt for allment tilgjengelege dokument (pli
Ambisjon for Digital Bevaring:
Mål for Digital Bevaring:
Roadmap
Veikart Digital bevaring 2024-2025
1. Overgang fra SAM-FS til DPS med HPSS som bit-repo
2. Flytte historisk materiale fra SAM-FS til DPS med HPS
repository (T)
Monthly delivery plan
2024 Februar leveringsplan Digital Bevaring
Main activities (epics)
DB-645 Techtalk: eArchiving (e-ark) som standardformat for bevaring
DB-651 Drift av DPS
DB-667 Finne ut hvordan vi kan motta flere filtre i utlevering, alik
DB-485 Endre parameter context til å bruke arv
DB-650 Test om Siegfried håndterer store filer bedre enn Droid
DB-682 Kan vi erstatte DroidIdentificationProcessor med Siegfrie
DB-666 Finne ut hvorfor Droid ikke identifiserer wav-filer riktig
DB-683 Oppdatere signaturfil for Droid
DB-613 Sende DPS logger til Logstash/ES/Kibana
DB-585 NB-pakker til E-ARK
DB-587 Reark DSM Radio
DB-607 Lage Pronom signatur for JSON Lines
DB-624 Rearkivering av historisk radiomateriale i DPS
DB-622 Teste rearkivering flyt ende til ende
DB-662 Finne ut om vi kan lage MD5 sjekksum for Representasjon
DB-665 Bytte ut Droid med Siegfried
DB-688 MP3 visningsfiler skal ikke flyttes fra DSM til DPS
Kanban board (tasks)
Digital Bevaring
Epic
Type
Custom filters
TIL UTFØRING 12
UNDER ARBEID 10
KLAR TIL VERIFIS
OCFL: Følge opp forslag fra Jürgen Enge
DIVERSE
DB-556
DB-622
DB-614
Endre parameter context til å bruke arv
DRIFT AV DPS
DB-488
Finne ut om vi kan lage MD5 sjekksum for Representasjons METS
REARK DSM RADIO
DB-662
Dokumentere hvilke verktøy som er aktuelle å bruke for validering og metadata extraction
DIVERSE
DB-454
Gjøre tester av ravecocked for encoding av DPX-sekvenser
DB-689
Blogpost Strategi
STRATEGI DIGREV
DB-686
Test om Siegfried håndterer store filer bedre enn Droid
DRIFT AV DPS
DB-650
Validere E-ARK sip
REARK DSM RADIO

BUILDING DOMAIN EXPERTISE

- Share experiences and policy documents on our blog
- Revision preferred file format list

- Membership **Digital Preservation Coalition** (DPC)
  - DPC assessment tools (DPC-RAM, DPC-CAT)
  - DPC bitlist council

- Involvement in national and international community



BIT REPOSITORY REPLACEMENT

2020-2022

- SAM-FS EOL (2021)
- Tender (2021)→IBM **High Performance Storage System** (HPSS)
- Installed (2022)

CLOUD VS. IN-HOUSE?

- Cons of cloud:
  - Possible **performance** challenges when moving large data volumes
  - **Legal uncertainties** in relation storing cultural heritage materials at commercial vendors, potentially outside of Norwegian borders
  - The costs of **retrieving** large amount of data from cloud provider
  - **Lack of in-house experience** with cloud infrastructure at the time
  - Solid **in-house experience with self-hosting** (we were comfortable to keep on doing it)

HPSS (IBM)

High Performance Storage System

- Linux OS
- Block based storage
- Mix of different disk and tape technologies (no vendor lock-in)
  - **Disk:** Fujitsu, Huawei og Nexsan
  - **Tape:** LTO8 in 2 SL8500 (10k slot libraries)

SELECTION OF HPSS

High Performance Storage System

- Supports 3-2-1 (disk+tape+tape)
- Scales well
- No vendor lock-in (multi-vendor HW)
- Multilevel checksumming (blocks and files)
- Large user community
  - 30+ clients and 3+Exabytes stored in HPSS systems worldwide

HPSS IMPLEMENTATION

Window of opportunity

- **Installation** → Production lines still writing to SAM-FS

- **Move bits as-is from SAM-FS to HPSS?**
- **Opportunity to do things better!**

- Establish new ingest/preservation/dissemination methodology according to principles

OFF-THE-SHELF PRODUCTS?

2021-2022

- **Criteria:**
  - Handles **large data volume** and expected growth
  - Need for **automated processes**
  - Has **separated preservation and playback** in different solutions
  - **Standardized** and **open solutions**
- Surveyed the market (looked at Archivematica, Libnova, CSC, and more)
- **None of these fit our needs – challenges with:**
  - Scale in **data volume**
  - **Licensing** (often volume-based)
  - **Running environment** that does not fit in NLN architecture
  - Systems contained **functionality** that NLN did not request (viewing/playback)

DIY → DPS

“Digital Preservation Services”

- **Developed by digital preservation team (Jun 2022 → Dec 2022)**
- **Built after preservation principles**

- **DPS 1.0** = Unified ingest workflow to HPSS
  - **Checksums** for all files (and SIPs)
    - Stored in HPSS (along with files), in HPSS (DB2 database), and in DPS (locationDB)
  - Standardized **delivery format** (not package format)
  - **Asynchronous** communication
  - **Inventory database** with information on:
    - Content type, File types, Number of files in package, Location in bit repository, Events regarding ingest, Who delivered SIP

DPS DEVELOPMENT

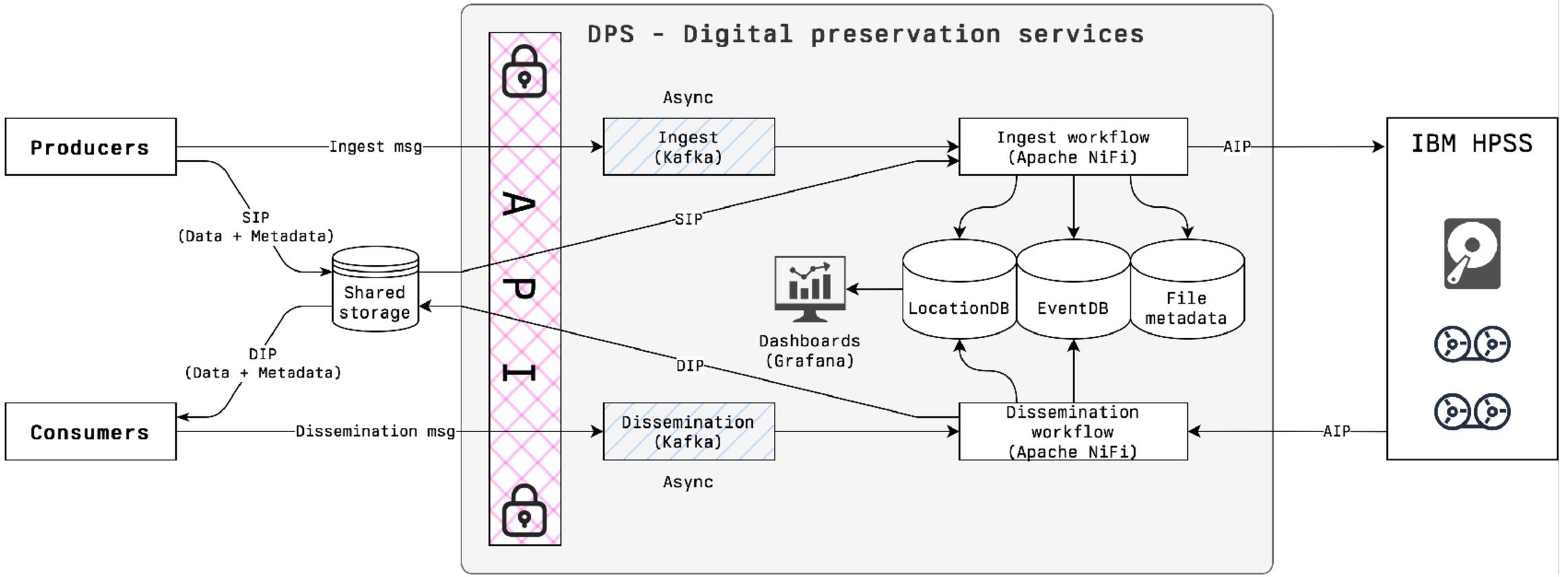
2022→2024

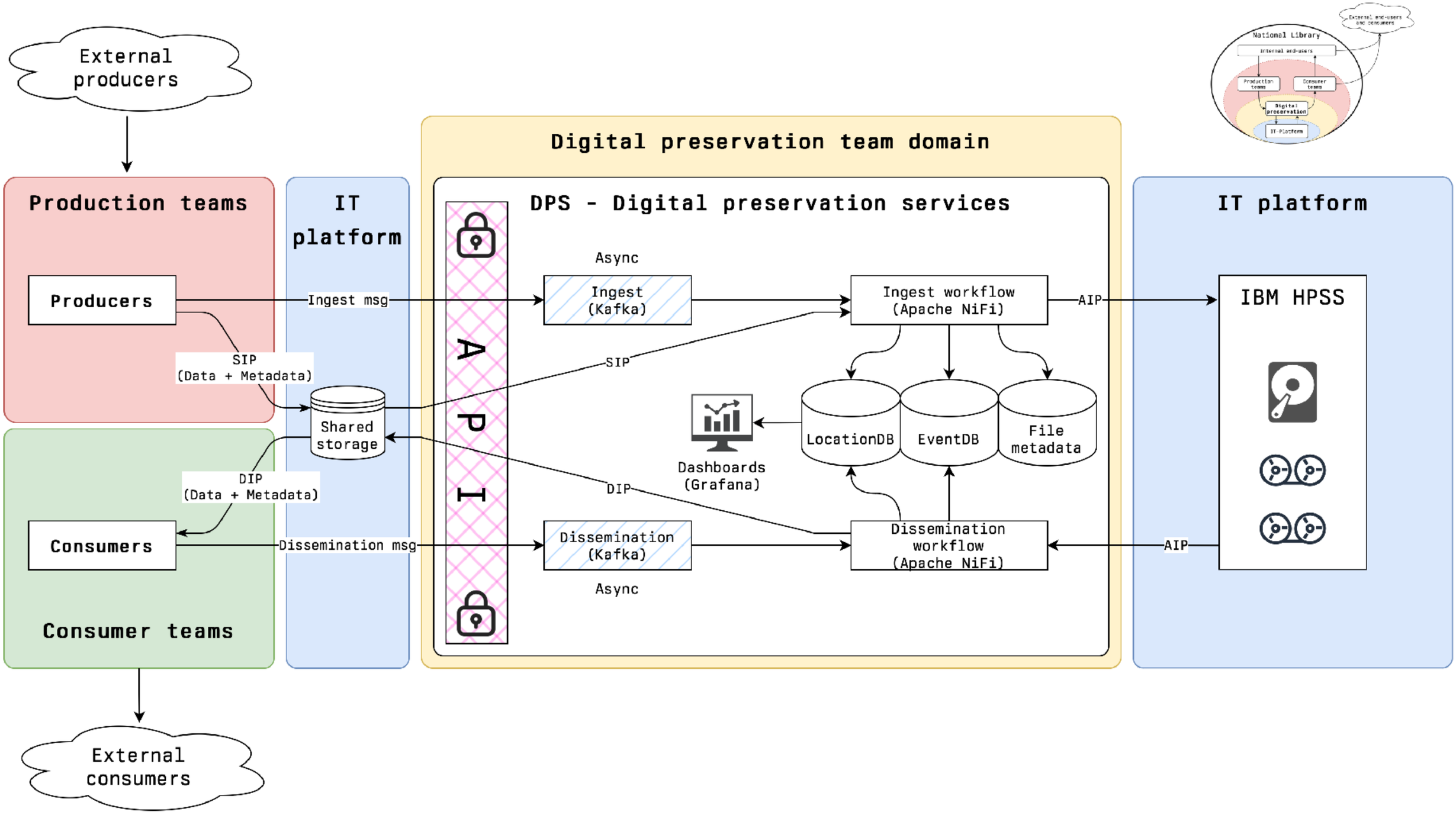
- **Iterative** and **incremental** development (MVP):
  - Addition of dissemination workflow (2023)
  - Ingest workflow expanded (2023→)
    - File identification (DROID/Siegfried)
    - File validation etc.

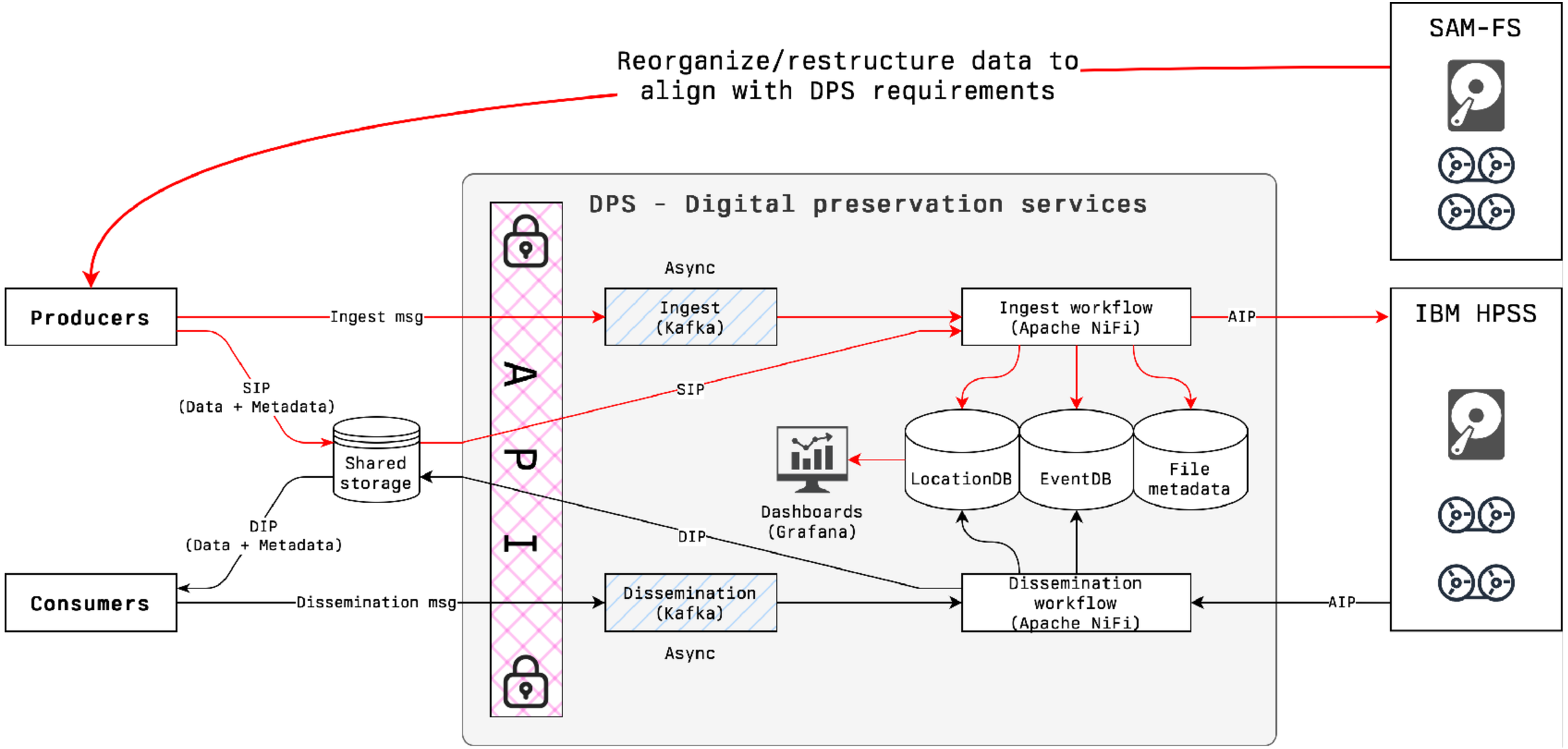
- Update 30+ production lines to deliver new data to DPS instead of SAM-FS.
  - Took most of 2023 to accomplish (1 production line remaining still!)

DPS TECHNOLOGIES

- **Java, Spring Boot, Keycloak, Kubernetes** for REST APIs for SIP+DIP messages
- **NFS/GlusterFS** shared storage for transfering SIP+DIP packages
- **Apache Kafka** for asynchronous transfer of messages
- **Apache NiFi** for processing SIP/AIP/DIP packages
- **IBM High Performance Storage System** (HPSS) for archival storage
- **Grafana** dashboards for monitoring and statistics


DPS - Digital preservation services
Producers
Consumers
Ingest msg
SIP
(Data + Metadata)
DIP
(Data + Metadata)
Dissemination msg
Shared storage
A P I
Async
Ingest (Kafka)
Dissemination (Kafka)
Async
SIP
DIP
Ingest workflow (Apache NiFi)
Dissemination workflow (Apache NiFi)
AIP
AIP
Dashboards (Grafana)
LocationDB
EventDB
File metadata
IBM HPSS

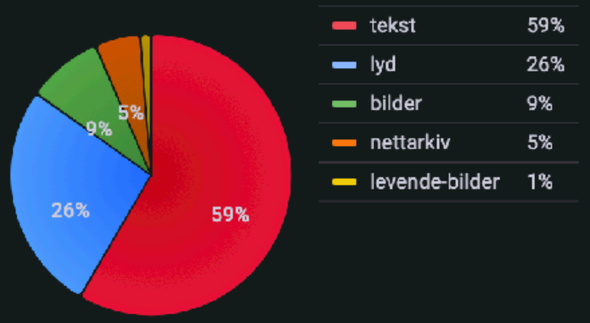

External producers
Production teams
Producers
IT platform
Ingest msg
SIP (Data + Metadata)
Shared storage
DIP (Data + Metadata)
Consumers
Dissemination msg
Consumer teams
External consumers
Digital preservation team domain
DPS - Digital preservation services
A P I
Async
Ingest (Kafka)
Ingest workflow (Apache NiFi)
AIP
SIP
DIP
Dashboards (Grafana)
LocationDB
EventDB
File metadata
Dissemination (Kafka)
Dissemination workflow (Apache NiFi)
Async
IT platform
IBM HPSS
AIP



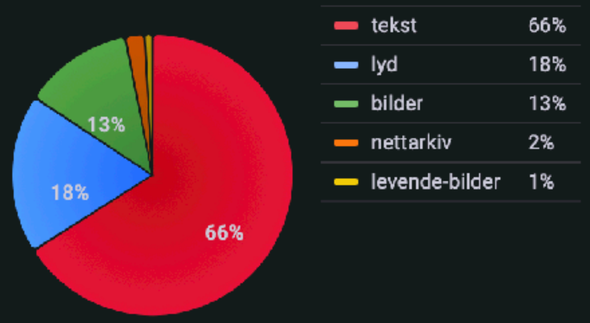
REARCHIVING SAM-FS→HPSS (2023→2025)

Number of AIPs and Files, and amount of Data by Category and Type

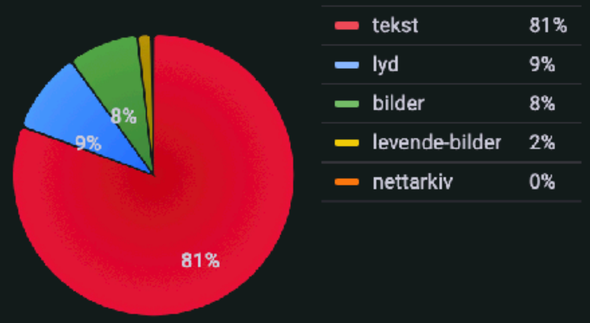
AIPs by Category



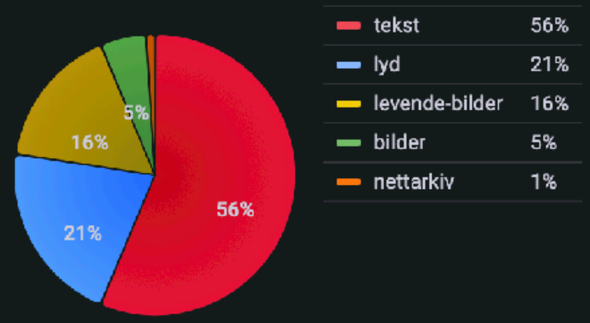
Files in HPSS by Category



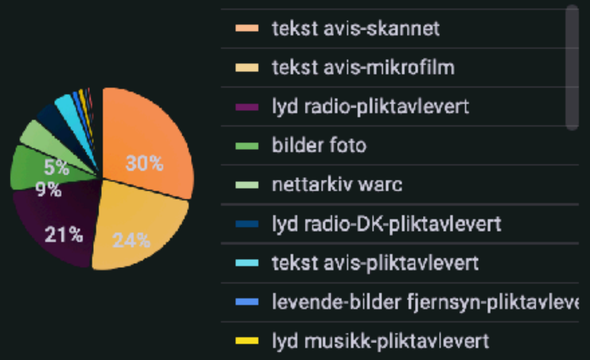
Single files preserved by Category



Amount of data in Bytes by Category



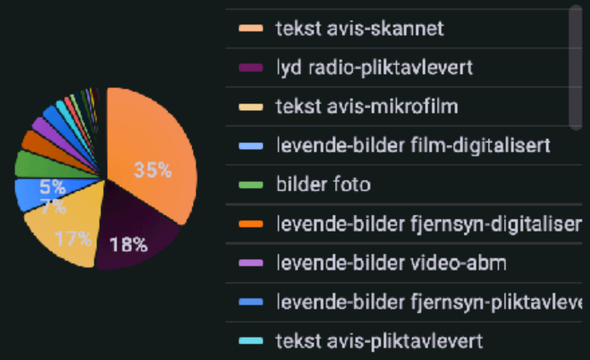
AIPs by Category and Type



Files in HPSS by Category and Type

Single files preserved by Category and Type

Amount of data in Bytes by Category and Type



AIPs by Category and Type

Category	Type	Count ↓
tekst	avis-skannet	2,382,151
tekst	avis-mikrofilm	1,898,377
lyd	radio-pliktavlevert	1,686,904
bilder	foto	711,289
nettarkiv	warc	424,441

Files in HPSS by Category and Type

Category	Type	Count ↑
nettarkiv	acquisition	21
levende-bilder	videokunstarki...	1,260
lyd	musikk-studio-f...	4,260
levende-bilder	film-digitalisert	5,227
levende-bilder	fjernsyn-DK-pli...	6,816

Single files preserved by Category and Type

Category	Type	Count ↓
tekst	avis-mikrofilm	104,184,196
tekst	tidsskrift-skannet	47,211,872
tekst	avis-pliktavlevert	46,641,050
bilder	foto	24,115,526
tekst	avis-skannet	22,608,577

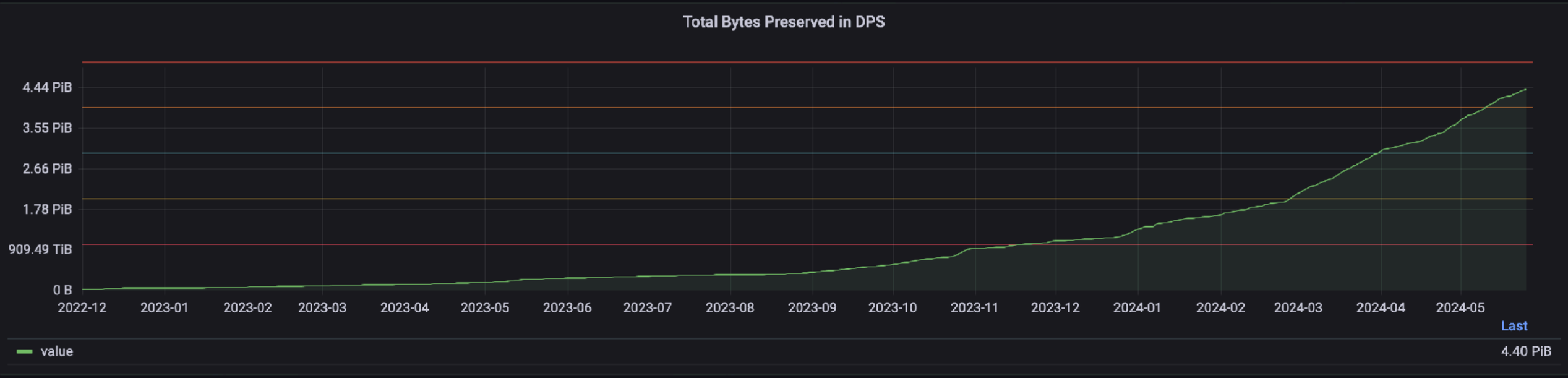
Amount of data by Category and Type

Category	Type	Bytes ↓
tekst	avis-skannet	1.50 PiB
lyd	radio-pliktavlevert	804.87 TiB
tekst	avis-mikrofilm	745.98 TiB
levende-bilder	film-digitalisert	288.14 TiB
bilder	foto	233.07 TiB

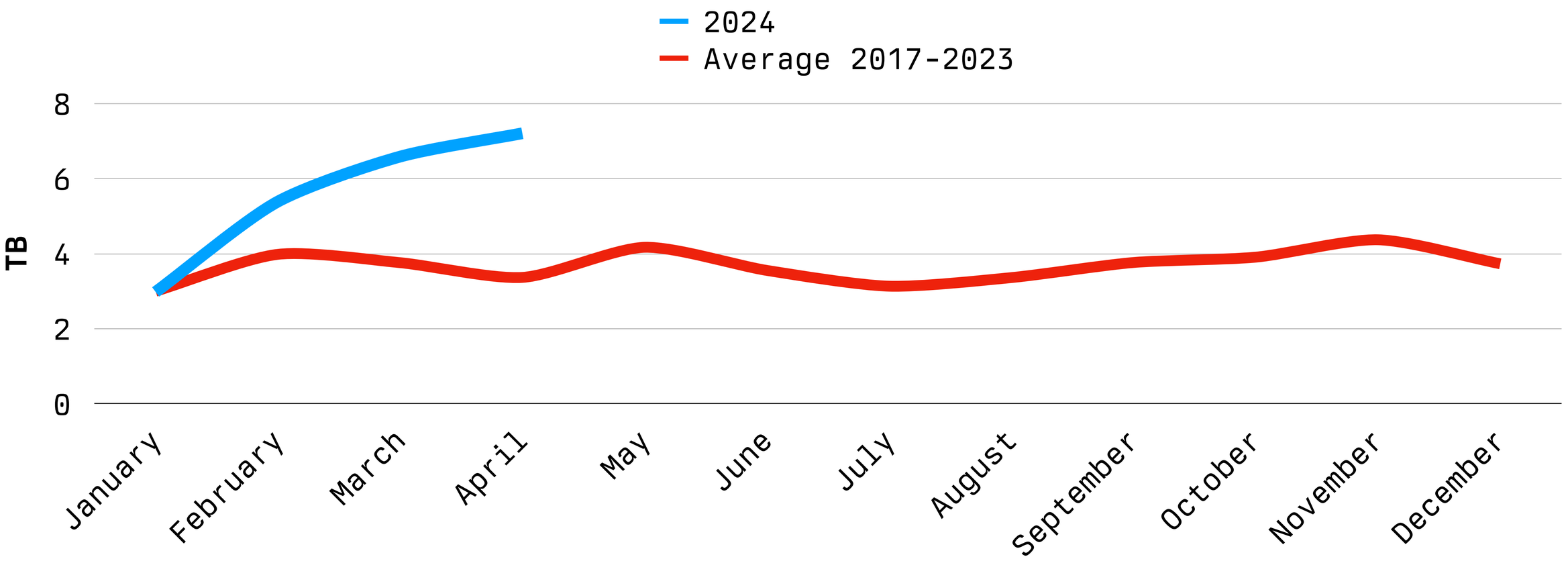
Total number of AIPs and Files, and amount of Data preserved

Total number of AIPs preserved	Total number of files stored in HPSS	Total single files preserved	Total amount of data preserved
8,210,715	39,450,431	298,141,949	4.42 PiB

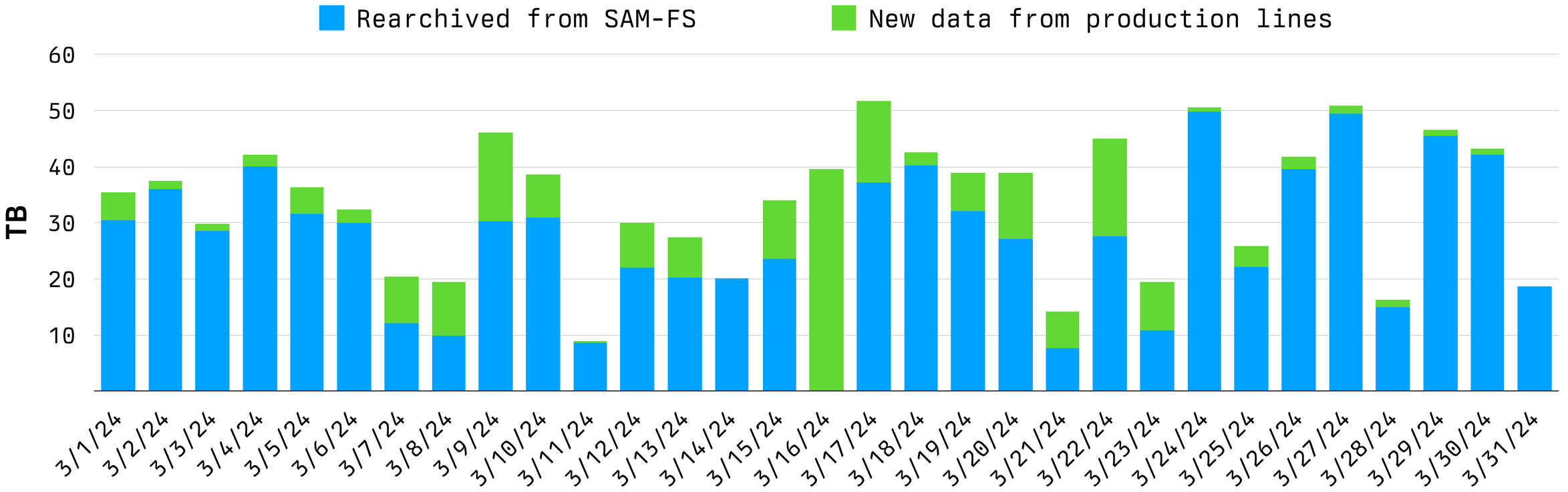
Total bytes preserved in DPS



AVERAGE DAILY INGEST OF NEW DATA PER MONTH IN TB



DAILY INGEST TO DPS IN TB, MARCH 2024



Ingest in TB In March	Total	Re-archived from SAM-FS	New data
Accumulated	1 035	838	197
Daily average	33	27	6,5

DPS 2.0 - FUTURE PLANS

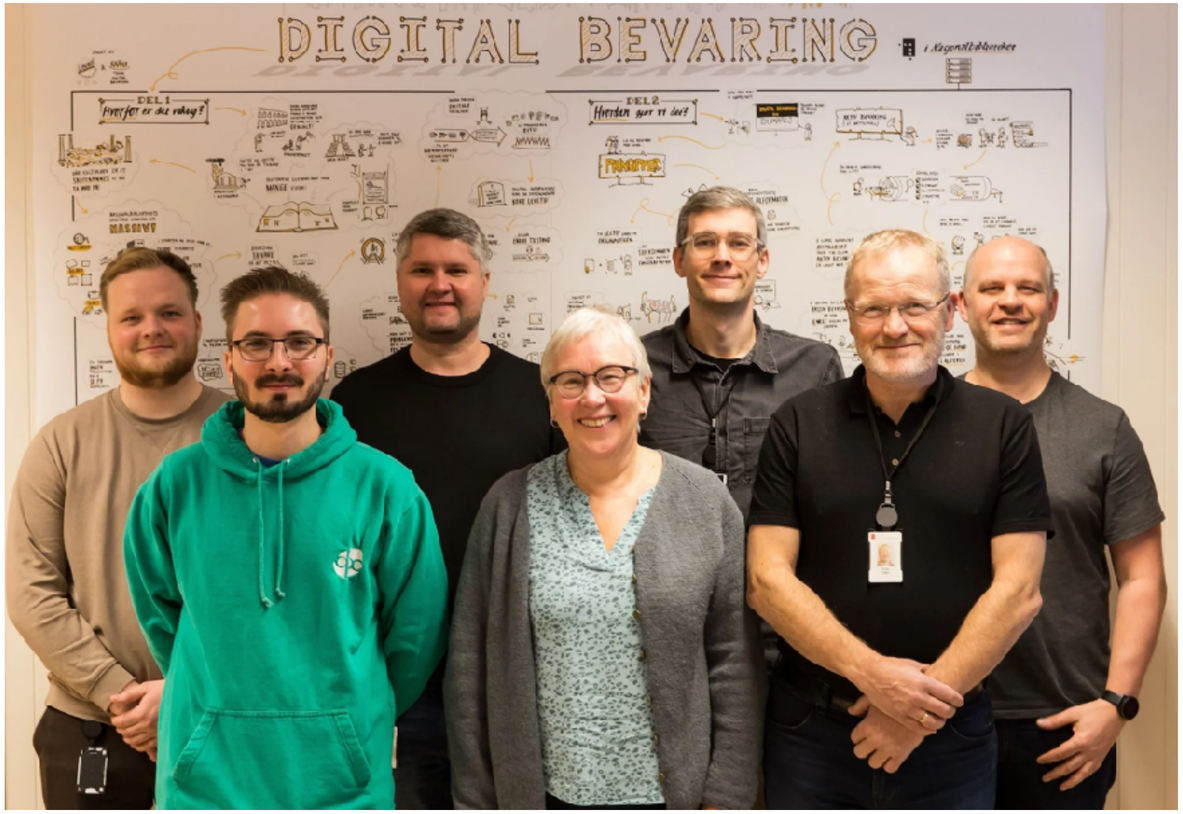
2024→2025

- **Keep improving in small increments!**

- **Authentication and Authorization**
  - Lock down the DPS (WIP)
  - Role-based access (near future)
- **Standardize information package content structure**
  - Implement eArchiving Standards & Specifications (WIP)
- **Improve ingest workflow to handle unpacked files**
    - Get control at file level (not .tar level)

CONTACT INFO

- torbjorn.pedersen@nb.no
- digitalpreservation-blog.nb.no/
- NB.no/



The digital preservation team