Datamuseum.dk

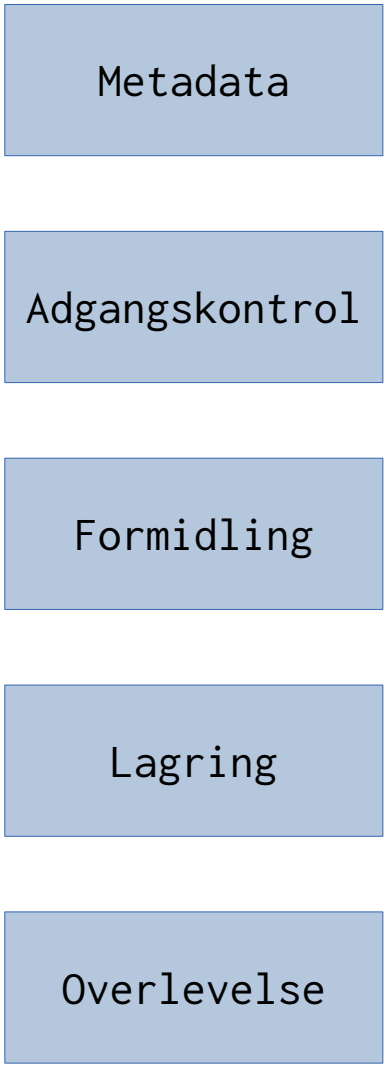
Bitarkivet - status 2020

Poul-Henning Kamp

phk@FreeBSD.org

phk@Varnish.org

@bsdphk



De nødvendige metadata
Format, Indhold osv.

GDPR, båndlagte donationer, almindelig anstændighed
Public, Private, Restricted, (Gone)

Dataformater
Integration med wiki mv.

WARC + index
AardWARC

Backup osv.

Hvad er Metadata ?

Metadata skal gøre det muligt at genfinde data

Metadata skal være søgbare

Metadata skal besvare spørgsmål:

- Hvad er det ?
- Hvor kommer det fra ?
- Hvor hører det til ?
- Hvem har adgang ?
- Hvor kan det genfindes ?
- Hvilket Format ?
- Hvad ved vi ellers ?

Metadata er fakta, ikke holdninger eller historier.

Metadata er dynamiske

Metadata kan forandre sig over tid:

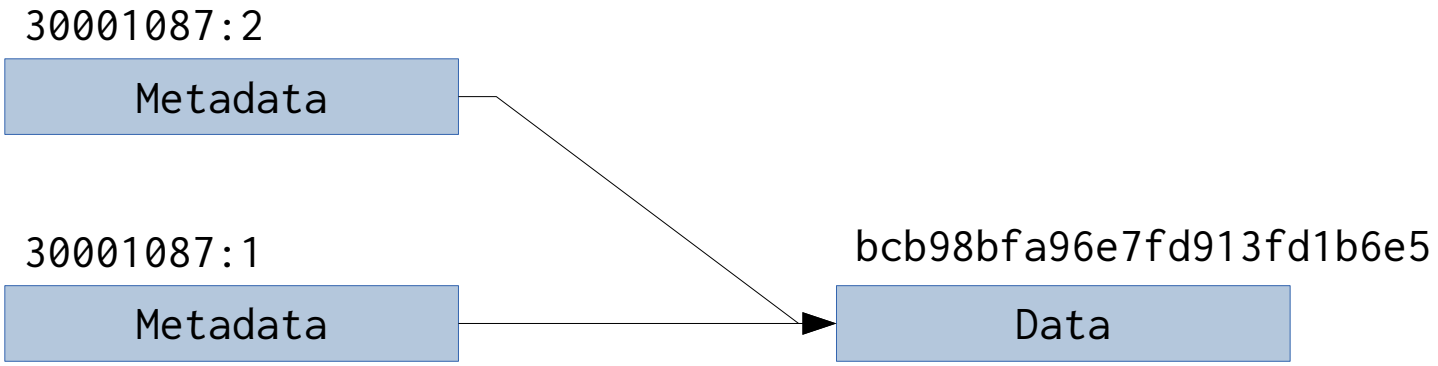
- Ændret adgangskontrol - åbne/lukke
- Ændret placering - (pt. ikke relevant)
- Mere korrekt viden - ”4 august 1975” frem for ”1975”
- Formatkonvertering - (helst ikke)
- Ændret objekt - f.eks med færre læsefejl

Vores lager understøtter ikke editering & sletning

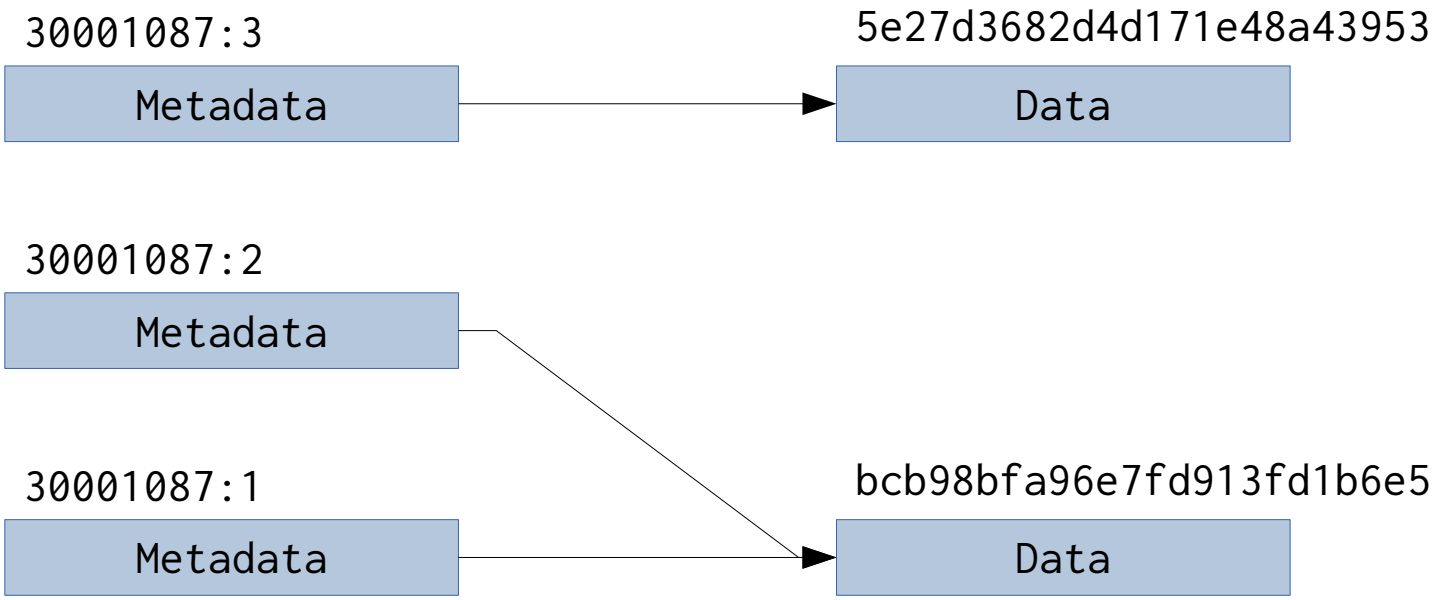
- Ergo: Versionering

Data & Metadata - Datamodel

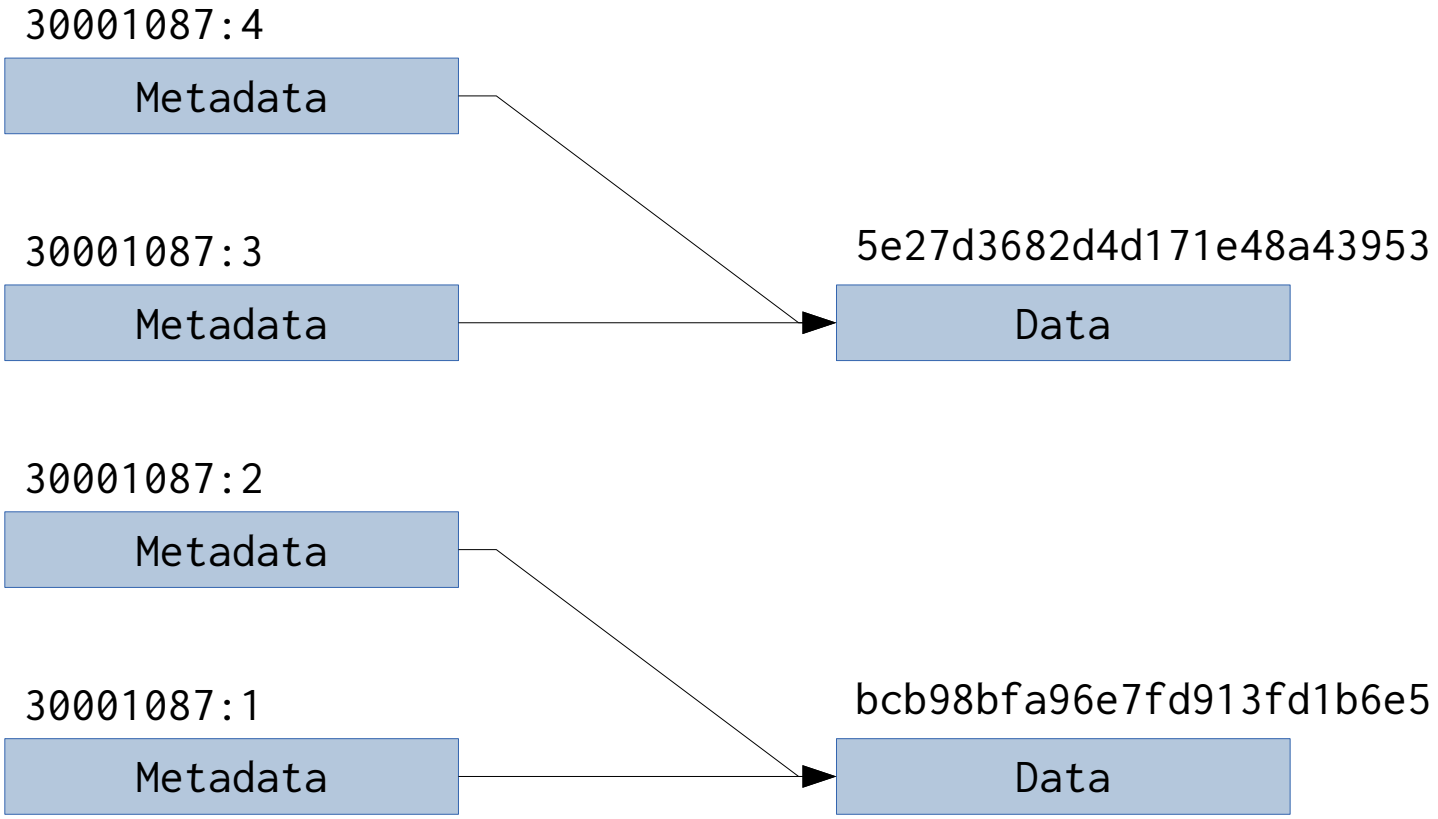
Data & Metadata



Data & Metadata



Data & Metadata



Metadata - Indhold

Obligatoriske afsnit:

BitStore.*	- Lagring og Adgang til objektet
DDHF.*	- Forbindelsen til samlingen

Søgefoder afsnit:

Document.*	- Dokumenter, tekstfiler
Image.*	- Fotos, grafik, tegninger, diagrammer
Media.*	- Diske(tter), bånd, strimler, hulkort
Presentation.*	- Video, Audio, Slides, Referater

Metadata - Format

Se: https://datamuseum.dk/wiki/Bits:Metadata

UTF-8 tekstfil med brutal syntax:

```
    * (
        Afsnit '.' felt ':' NL
        * ( TAB feltindhold NL )
        NL
      )
    '*' 'E' 'N' 'D' '*' NL
```

Hvert afsnit og felt har specifik definition:

- Obligatoriske/Valgfri felter
- En/flere linier
- Valgmuligheder for værdi
- Foreskrevet format
- Konsistenscheck (f.eks RCSL & ISBN)

} Python3-kode kan anvendes UX/validaring

Metadata - Eksempel 1/2

```
BitStore.Metadata_version:
        1.0

BitStore.Access:
        public

BitStore.Filename:
        BIPAR_01.TXT

BitStore.Size:
        2530

BitStore.Format:
        ASCII

BitStore.Ident:
        30001087:1

BitStore.Digest:
        sha256:5e27d3682d4d171e48a439539fbf86de9424a7f264edd357c468c52d5a0d6b97

BitStore.Last_edit:
        20200106 phk
```

Metadata - Eksempel 2/2

```
DDHF.Keyword:
        RCSL/43/GL
        RC3600/DOMUS

Document.Title:
        RCSL-43-GL-5317 BIPAR.01 Source code

Document.Author:
        MLM

Document.Date:
        77.10.20

*END*
```

Metadata - DDHF.Keyword

Træstruktur - primært til indexering på wiki'en

B&O/{DOCS|SW}
CBM900
COMAL/{RC3600|RC700|Z80}
DDE/SUPERMAX
DKUUG/{EUUG|DKUUGNYT}
FACIT/{ABC80|TWIST}
GIER/{ALGOL_II|ALGOL_III|ALGOL_4|…}
ICL/COMET
PÆDAT
RC/RCNET
RC2000
RC3500
RC3600/{COMAL|DE2|DOMUS|HW|LOADER|MUSIL|…}
RC4000/{HW|TEST|SW|MATHEMATICS}
RC7000
RC8000
RC850
RATIONAL_1000/{DISK|TAPE|SW|DOCS}
SCMETRIC/MICROMUX
SINCLAIR/ZX80

Adgangskontrol

BitStore.Access: {public|private|restricted|gone}

public:
    Fri adgang for alle
    Udgangspunktet

private:
    Adgang med wiki-konto = Medlemmer
    Personlige beretninger, almindelig anstændighed

restricted:
    Adgang kun efter aftale med formand/bestyrelse
    GDPR, Straffeloven, almindelig anstændighed

gone:
    Udgået af samlingen

Adgangskontrol

Bitstore tilbyder kun to web-services for omverdenen:

Get-Object: http://bits.ddhf.dk/bits/30000100

Get-Metadata: http://bits.ddhf.dk/meta/30000100

Beggge checker Wiki-cookie for adgang til ”private”

Begge fejler på ”gone” og ”restricted”

Bitstore pusher index+pages til Wiki

For ”public” lagres metadata i wiki-page
= søgbart med wiki’ens søgefunktion

Andre → wiki-page bruger template og (privat) API.

Formidling - presentation

Alle objekter undtagen ”gone” spejles ind i wiki’en

DDHF.Keyword er primær søge/browse adgang:

```
DDHF.Keyword:
    COMAL/RC3600
    RC3600/COMAL
    RCSL/43/GL
```

DDHF.Genstand linker til registreringssystemet

Direkte links i Wiki-sider:

```
[[Bits:30000916|Rational 1000, Guru Course 01]]
```

Formidling - Formater

Det lange perspektiv styrer: Skal kunne læses om 100 år

- Ingen formater der kun kan læses med bestemt software
- Ingen kryptering og kodeord
- Ingen Turing komplette formater
- Tabsfrie formater
- Formater der kan genskabe originalen
- Ikke flere formater end nødvendigt
- Formater der er så brugbare som muligt

Formidling - Formater

Udgangspunkter:

Dokumenter:	PDF, ASCII{_EVEN|_ODD}, GIERTEXT, EBCDIC…
Billeder:	PNG, TIFF, JPEG
Hulstrimmel:	GIERTEXT, ASCII{…}, EBCDIC, BINARY, …
Bånd:	SIMH-TAP (tape-marks, blocksize), BINARY
Diske(tter):	BINARY (Array af sektorer) KRYOFLUX hvor nødvendigt (ex: CBM900)
Video:	MPEG (Hvilke Codecs ?!)
Hulkort:	Som Hulstrimmel

Formidling - Formater - Brugbarhedskompromisser

Hulstrimmel: Rå binær kopi
   + Strimlen kan genskabes
   - Browsere kan ikke ASCII med paritet eller GIERTEXT

DDE diskimages: Rå binær kopi
   → Vi kender ikke formatet godt nok til at skille skæg fra snot (Oracle, Veritas mv.)

PC diske: Uddrag relevante filer
   + Win+Office ofte ≥ end det interessante
   + Gemmer ikke ”tom plads”
   - Skæg/Snot stillingtagen nødvendig

Regulære bånd, (tar, cpio, dump): Rå binær kopi

Iregulære bånd, (boot, multi-volume etc): SIMH-TAP

Formidling - Formater - Bundtning

Som udgangspunkt gemmes objekter enkeltvis

Flere objekter kan bundtes, hvis det er mere brugbart

Eks på bundtning:
- Kildetekst spredt over mange filer
- Kryoflux Stream-filer
- Hjemmekatalog

Eks på ikke-bundtning:
- Kildetekst i få filer
- Floppy 1-3 til CBM900
- Disk 1-10 til DDE SuperMax
- Katalog med PDF filer

Til bundtning bruges altid:ZIP(BAGIT(SHA256))

Formidling - Formater

Objekter tilbydes på browser-venlig vis:

- BitStore.Filename → Forslag til filnavn ved download
- BitStore.Format → MIME type så f.eks billeder vises
- ”application/binary” når browsere ikke fatter formatet

Planlagt ”efterbrænder” til bedre presentation

- GIERTEXT/ASCII_EVEN/EBCDIC → UTF8 konvertering
- Hexdump m/ paginering
- Browsernavigation af filsystemer, BAGIT bundter osv.

Lagring - WARC - ISO-28500

WARC - ”Web ARChive format”
   Opfundet af Archive.Org, DKB meget involveret

Filformat: * (GZIP(WARC-header) GZIP(WARC-body))

Append only

Support for segmentering af store objekter

Alle records har en unik id:

```
WARC-Record-ID: <http://datamuseum.dk/bits/81c3938000000080000000>
```

Alle records har checksum:

```
WARC-Block-Digest: sha256:ef8c6dff31114b8818a203e505d29214c4766968faf5488ff3b9d3
```

Lagring - WARC - ISO-28500

```
WARC/1.1
WARC-Record-ID: <http://datamuseum.dk/bits/81c393800000000080000000>
Content-Length: 539
Content-Type: text/plain;charset=utf-8
WARC-Block-Digest: sha256:ef8c6dff31114b8818a203e505d29214c4766968faf5488ff3b9d367aa
WARC-Date: 2019-06-10T17:43:58Z
WARC-Refers-To: <http://datamuseum.dk/bits/5ab8ba1f9be32a04a6431189>
WARC-Type: metadata

BitStore.Metadata_version:
        1.0

BitStore.Access:
        public
[…]
```

Lagring - AardWARC

https://github.com/bsdphk/AardWARC

Get, Put, Indexering, Audit, Rebuild osv.

Warc-ID	TYPE	SILO	OFFSET	NEXT
05c393800000000080000000	0x00000008	0	97985436	00000000
05d24e8762e7dd8929c99d62	0x00000004	1	845078614	00000000
05e393800000000080000000	0x00000008	2	1073696557	00000000
05edb0208d9d51fdde7e1458	0x00000004	1	939080879	00000000
05f0d65bc02d4a9c7295fc31	0x00000004	0	328949166	00000000
05f2873c798715988ef3f2f0	0x00000010	43	660	f43d6272
05f58f8d513c3a66299ff7de	0x00000004	0	637216573	00000000
06217c2911b9cb0cb3112853	0x00000004	0	375873790	00000000
062393800000000080000000	0x00000008	0	637154440	00000000

Warc-ID er velfordelt = direkte addresseret index:

$0x05f0d65b / 2^{32} = 0.02320$

Kan findes ca. 2.3% inde i sorteret indexfil

Lagring - Warc-ID’s

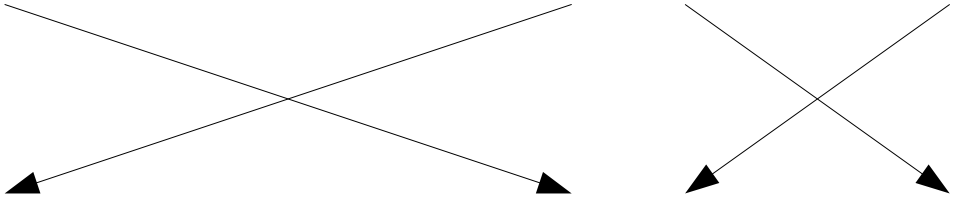
Objekter: Første 96 bit af SHA256 sum

Hurtigt at checke om objekt allerede findes

Metadata bruger venstrekørsel (spredning i index):

0x04a393800000000040000000

0x04a3938000000000:0x40000000



0x0000000001c9c520:0x00000002

30000416:2

Overlevelse

In-House Backup (ukrypteret)

- På server S1 på Tapeten

Off-site Backups (krypteret, hos betroede medlemmer)

- Prioritering:
  - Beskyttelse imod tyveri & kassation af datamedier
  - Værten beskyttet imod kendskab til indholdet
  - Overlevelse af mediefejl
  - Rsync-venligt
  - Båndbredde-begrænsning
  - Validering

- AES256CTR-then-HMAC (Se Første Colin-bog)

- https://datamuseum.dk/wiki/Privat:BitStoreBackup

Overlevelsemuligheder

Det Nationale Bitmagasin:
   Oprettelse: 10000,-
   Per år: 10000,- (ca pris for 2TB)

TarSnap.com:
   Kompetent krypteret backup-service
   250 picodollars per byte-month ~= $6000/år

Hvis alt andet svigter:

   WARC filer efterlades på DKB’s dørtrin en mørk nat

Mangler:

Automatiseret Workflow

- Leder efter en Round Tuit

Webform til indtastning af metadata + upload af objekt

- Hjælp søges!

Presentations-efterbrænderen

- Hjælp søges!

Udadvendt presentation af fotos/billeder med BA backend

- Hjælp er på vej (?)