Filorganisering og filformat

RDA-kurs i FAIR forskingsdatakuratering
UiT, 20.-21. januar 2020

Philipp Conzett, UiT

I går...

- Generelle metadata >> gjenfinnbar; jf. F-en i **F**AIR
  (særleg viktig: Title, Description, Keywords)
- Metadatastandardar >> interoperabel; jf. I-en i FA**I**R
  (Ikkje nemnt i går: *De jure*- vs. *de facto*-standardar)
- Dokumentasjon >> gjenbrukbar; jf. R-en i FAI**R**

Gode rutinar for **filorganisering og filformat** er også med på å moglegjera **gjenbruk** av data.

Filorganisering

- God **filorganisering** består av to ting: **gode filnamn** og **god mappestruktur**.
- Korfor er det viktig med gode filnamn?
  >> Gjer det mogleg/lettare å **identifisera og (gjen)bruka** filer på ein effektiv måte; jf. R-en i FAIR.
- To viktige innfallsvinklar til **namngjevingsstrategiar**:
  1. Gode filnamn indikerer **innhaldet** og **versjonen** av ei fil, og tener som grunnlag for å **identifisera**, **klassifisera** og **sortera** filer.
  2. Namngjevingsstrategiar er **konsistente** over tid og mellom personar som er involverte i datahandteringa.

Kjelde: CESSDA Training Working Group (2017 - 2019). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Henta frå https://www.cessda.eu/DMGuide

Vanlege element i eit filnamn

- Versjonsnummer
- Dato for oppretting
- Namn på den som har laga fila
- Beskriving av innhaldet
- Namn på forskargruppe, institutt eller organisasjon som har med dataa å gjera
- Dato for publisering/tilgjengelegjering
- Prosjektnummer

Kjelde: CESSDA Training Working Group (2017 - 2019). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Henta frå https://www.cessda.eu/DMGuide

Gode råd for korleis filnamn skal skrivast

- Bruk meiningsfylte/deskriptive, men korte namn.
- Bruk filnamn som klassifiserer filtypane: `interview, transcript, notes`
- Unngå mellomrom: `Project notes.pdf`
  Bruk understrek, bindestrek eller pukkelord (“camel case”) i staden:
  `Project_notes.pdf | Project-notes.pdf | ProjectNotes.pdf`
- Unngå spesialteikn, inkl. skandinaviske bokstavar:
  `“ / \ : * . ? ‘ < > [ ] ( ) & $ æÆ øØ åÅ`
- Bruk internasjonalt format på datoar: `ÅÅÅÅ-MM-DD`
- Ikkje bruk filekstensjonar i sjølve filnamnet: `Project_notes_pdf.pdf`
- Inkluder versjonsnummer dersom nødvendig: `interview_v_03.pdf`

Kjelde: CESSDA Training Working Group (2017 - 2019). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Henta frå https://www.cessda.eu/DMGuide

Filsortering

Gode filnamn mogleggjer filsortering på ulike måtar:

Sortert etter dato:
1955-04-12_notes_MassObs.docx
1955-04-12_questionnaire_MassObs
1963-12-15_notes_Gorer.docx
1963-12-15_questionnaire_Gorer.pd

…innhald:
Gorer_notes_1963-12-15.docx
Gorer_questionnaire_1963-12-15.pdf
MassObs_notes_1955-04-12.docx
MassObs_questionnaire_1955-04-12.pdf

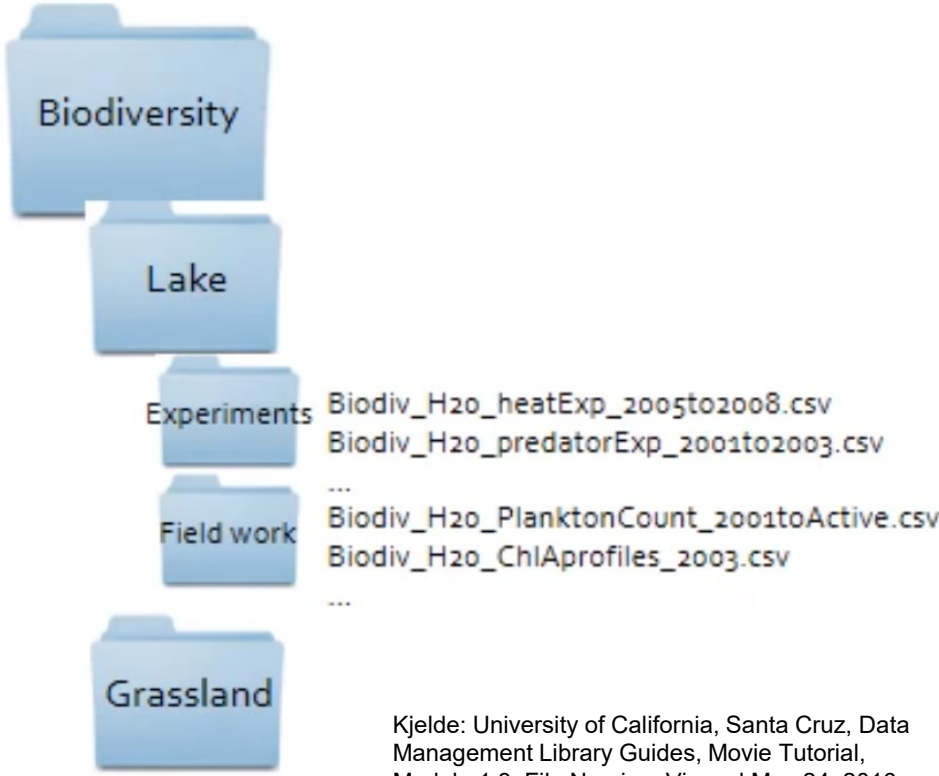
…filtype:
Notes_Gorer_1963-12-15.docx
Notes_MassObs_1955-04-12.docx
Questionnaire_Gorer_1963-12-15.pdf
Questionnaire_MassObs_1955-04-12.pd

Tvinga sortering med nummerering:
01_MassObs_questionnaire_1955-04-12.pdf
02_MassObs_notes_1955-04-12.docx
03_Gorer_questionnaire_1963-12-15.pdf
04_Gorer_notes_1963-12-15.docx

Kjelde: The University of California, Santa Cruz, Data Management Library Guides, Movie Tutorials, Module 1.3: File Naming. Viewed May 24, 2016 at http://guides.library.ucsc.edu/datamanagement/movies

Fil- og mappeorganisering

- Vel eit konsistent oppsett også for mapper.
- Hovudstrukturen i mappeorganiseringa bør også koma til uttrykk i filnamna. Viktig i tilfellet mappestruktur ikkje er støtta ved arkivering eller gjenbruk.
- Dokumenter fil- og mappeorganisering i ReadMe-fila.



Kjelde: University of California, Santa Cruz, Data Management Library Guides, Movie Tutorial, Module 1.3: File Naming. Viewed May 24, 2016 at https://guides.library.ucsc.edu/datamanagement/movies

Filformat

Korfor er **val av filformat** viktig for **gjenbruk** av data, og då først og fremst gjenbruk **på lang sikt**? (Jamfør R-en i FAIR)

- **Kompatibilitet:** Ikkje gjeve at ei fil som er laga i ein tidlegare versjon av ei programvare, framleis kan opnast og lesast (utan problem) i den aktuelle versjonen av programvara.
- Kanskje det **ikkje** kjem **nye versjonar** av programvara i det heile teke.
- Kanskje det vil vera mogleg å konvertera fila til eit nytt format, men slik **konvertering** kan innebera **reduksjon i kvalitet** eller jamvel **tap av data**.

Kjelde: CESSDA Training Working Group (2017 - 2019). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Henta frå https://www.cessda.eu/DMGuide

Arkivverdige eller føretrekte filformat

For å sikra gjenbruk på lang sikt bør data lagrast i eit arkivverdig filformat, dvs. eit format som eignar til langtidsbevaring. Slike format er vanlegvis

- ikkje-proprietære,
- opne, og følgjer dokumenterte internasjonale standardar,
- bruker standard teiknkoding (t.d. ASCII, UTF-8), og
- ikkje komprimerte.

Arkivverdige filformat for vanlege dokumenttypar

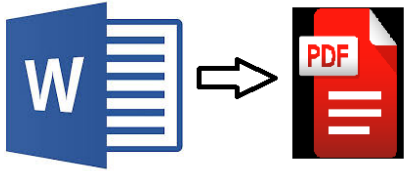
Dokumenttype	Ikkje-arkivverdig format (eksempel)	Arkivverdig format
Formatert tekst	MS Word (.docx)	PDF/A
Rekneark	MS Excel (.xlsx)	Tabulatorseparert Unicode-UTF-8-tekst (.txt)
Bilete	Windows Bitmap (.bmp)	Ukomprimert TIFF
Lyd	AAC (.m4a)	WAV
Video	Quicktime (.mov)	MPEG-4
Database	MS Access (.accdb)	XML eller tabulatorseparert Unicode-tekst (.txt)

Formatert tekst >> PDF/A

Hvis formatering av tekst er nødvendig (viktige lineskift, tabulatorar, figurar), bruk PDF/A. PDF/A er ein arkivverdig variant av PDF som m.a. har desse eigenskapane:

- Ikkje tillate med integrert
  - audio
  - video
  - køyrbart innhald (“executable content”)
  - eksterne ressursar
  - kryptering
- Grafiske element og fontar må vera integrerte i fila.
- Fila fungerer uavhengig av plattform og operativsystem.

Demo:
.docx >> PDF/A



Rekneark/tabelldata >> rein tekst

- Ulike gjengse format: .xlsx (Microsoft Excel), .ods (LibreOffice Calc), osb.
- Kan konverterast til rein tekstformat i ulike variantar:
  - kommaseparert (.csv) (i Noreg også semikolonseparert, ettersom komma er brukt som desimalteikn)
  - tabulatorseparert (.txt eller .tsv eller .tab)
- Separator i cellene kan vera problematisk ved importering til rekneark eller ved bruk i anna programvare. Tabulatorseparert rein tekst er derfor truleg mest problemfritt, for som regel er det ingen tabulotorar i cellene.
- Bruk standard teiknkoding, t.d. ASCII eller Unicode UTF-8.
- Kvart ark må konverterast.
- Figurar og diagram? Konvertér til PDF/A.

Demo:
.xlsx >> rein tekst
- i Excel
- i NotePad++

(Meir om teiknkoding)

- **ASCII** går greitt for vanlege teikn (tal, komma, punktum osb., latinske bokstavar utan aksent, ingen æ,ø,å m.fl.).
- For andre teikn, inkl. teikn frå andre alfabet, bruk **Unicode UTF-8** (utan BOM).
- Om **BOM**:
  - The UTF-8 BOM is a sequence of Bytes at the start of a text-stream (0xEF,0xBB,0xBF) that allows the reader to more reliably guess a file as being encoded in UTF-8.
  - Normally, the BOM is used to signal the endianness [byterekkjefølgja] of an encoding, but since endianness is irrelevant to UTF-8, the BOM is unnecessary.
  - According to the Unicode standard, the BOM for UTF-8 files is not recommended. Det kan føre til problemer med kommandofiler og programspråksfiler som ikke forstår hva en BOM er.

(Kjelde: https://stackoverflow.com/questions/2223882/whats-the-difference-between-utf-8-and-utf-8-without-bom)

Tap eller endring av informasjon under konvertering

- Rekneark:
  - Formatering, t.d. fargekodar
  - Kommentarar
  - Datoformat til reint talformat, t.d. `21.01.2010` >> `40199,00`
- Bilete
  - Redusert oppløysing
  - Færre fargar
- Lyd
  - Redusert lydkvalitet

>> Konvertering bør utførast og kvalitetssjekkast av nokon som er godt kjend med dataa.

Kor finn ein informasjon om arkivverdige format?

- Arkiv og bibliotek har retningsliner for føretrekte filformat, t.d.
  - Library of Congress
  - UK Data Service
  - UK National Archive (PRONOM)
  - DANS
  - DataverseNO
- Forskaren bør gjera seg tidleg kjend med retningslinene til arkivet han/ho skal bruka for å unngå ekstra arbeid før arkivering.
- Arbeid på gang med internasjonale tilrådingar for arkivverdige filformat
  - GO FAIR Implementation Networks og Convergence Matrix, samla/koordinerte tilrådingar for korleis FAIR kan implementerast på ulike fagområde
- Spør i RDA-interesse og -arbeidsgrupper!

Filformat og filekstensjon

- Filformat er uavhengig av filekstensjon / ekstensjonen bestemmer ikkje formatet / eitt og same format kan ha ulike ekstensjonar. Døme:
- Rein tekst:
  - .txt
  - .csv
  - .tsv, .tab
  - .R
  - .info
  - ...

Neste post på program

Gruppearbeid

>> Sjå delt Google-mappe.