

# Tutkimusaineiston hallinta

---

*Annaleena Okuloff*

*Terveystieteiden tieteenala-asiantuntija*

*3.5.2016*



TIETOARKISTO



# Tutkimusaineiston elinkaari

- Suunnittelu
- Keräys
- Tarkistus
- Dokumentointi
- Analysointi
- Arkistointi



# Suunnittelu – lähtökohtana jatkokäyttö

- Suostumus jatkokäytön mahdollistavassa muodossa
- Tutkimussuunnitelmaan eettistä lausuntoa varten kirjaus aineiston arkistoinnista ja jatkokäytöstä
  - Ei hävityslupausta!
  - Tällä hetkellä eivät ole tottuneita jatkokäyttöön
- Suunnitelma, miten tutkittavien tietosuojasta huolehditaan jatkokäytössä



# Suunnittelu – lähtökohtana jatkokäyttö

- Työnantajan kanta
  - Selvitä voimassa oleva datapolitiikka – Useimmat yliopistot ja tutkimuslaitokset suosittelevat arkistointia
- Omistajuus ja käyttöoikeus
  - Aineiston omistaja tutkija vai työnantaja?
  - Entä jos aineiston omistaja vaihtaa työpaikkaa? Jatko-opiskelijoiden ja muiden tutkijoiden asema?
  - Kenellä on käyttöoikeus?



# Datan keräys

- Kyselylomakkeen + muun kerättävän tiedon suunnittelu
  - Huomio jo aineiston käyttöön
  - Mukaan lääkäri, statistikko, datamanageri, projektin vetäjä
- Analysointi suunnitellaan tässä vaiheessa
  - Data oikeassa muodossa
  - Kysymysten muotoilu oikein



# Kvantitatiivinen aineisto

- Havaintoyksikkö on useimmiten henkilö
- Kyselylomakkeen vastaukset sekä muut tiedot omiin muuttujiinsa
- Yhteen ajankohtaan liittyvät tiedot yhdelle riville, jos mittaus toistetaan, uudet tiedot uudelle riville



TIETOARK  
FINNISH S  
DATA ARC

VIEWTABLE: Work.Jalka

	Henkilo	Kysely	K1	K2	K3	K4	K5	K6	K7	K8
1	1	Aloitus	3	4	2	2	3	3	3	2
2	2	Aloitus	3	3	2	3	3	3	3	3
3	3	Aloitus	3	3	2	3	3	3	3	2
4	4	Aloitus	3	4	2	3	3	3	3	2
5	5	Aloitus	3	4	2	3	3	2	3	2
6	6	Aloitus	3	2	2	3	3	2	3	3
7	7	Aloitus	3	3	1	2	2	3	3	2
8	8	Aloitus	4	5	1	3	2	3	3	2
9	9	Aloitus	3	4	2	3	3	3	3	3
10	10	Aloitus	3	2	2	3	3	3	3	2
11	11	Aloitus	3	3	2	2	2	2	2	2
12	12	Aloitus	4	4	1	2	2	1	2	2
13	13	Aloitus	3	3	2	2	3	1	2	1
14	14	Aloitus	3	3	3	3	3	3	3	2
15	15	Aloitus	3	3	2	3	3	2	3	2
16	16	Aloitus	4	4	1	2	2	2	3	2
17	17	Aloitus	3	3	2	3	3	3	3	3
18	18	Aloitus	4	4	1	1	1	2	3	1
19	19	Aloitus	2	3	2	3	3	3	3	3
20	20	Aloitus	4	3	1	2	3	2	3	2
21	21	Aloitus	3	3	2	3	3	3	3	3
22	22	Aloitus	2	3	3	3	3	3	3	2
23	23	Aloitus	2	3	2	3	2	3	3	2
24	24	Aloitus	3	3	1	2	2	2	3	2
25	25	Aloitus	2	3	2	3	3	2	3	2
26	26	Aloitus	4	3	2	2	3	2	3	2
27	27	Aloitus	2	3	3	3	3	2	3	3
28	28	Aloitus	2	1	1	3	3	2	2	3
29	29	Aloitus	2	3	2	3	3	3	3	2
30	30	Aloitus	3	3	2	3	3	2	3	2
31	31	Aloitus	3	3	2	3	3	2	3	3
32	32	Aloitus	4	3	2	3	3	2	3	3
33	33	Aloitus	4	3	2	3	3	3	3	2
34	34	Aloitus	4	3	2	3	3	2	3	2
35	35	Aloitus	2	3	2	3	3	3	3	2
36	36	Aloitus	3	3	2	2	3	3	3	1
37	37	Aloitus	3	3	2	2	1	2	3	2
38	38	Aloitus	3	2	1	2	2	2	3	2
39	39	Aloitus	3	3	1	2	3	3	3	2
40	40	Aloitus	3	3	2	3	3	3	3	3
41	41	Aloitus	2	3	2	2	2	2	3	1
42	42	Aloitus	3	3	2	2	2	2	3	2
43	1	3kk	3	3	2	3	3	3	3	3
44	2	3kk	3	3	2	3	3	3	3	3
45	3	3kk	3	3	2	3	3	3	3	3
46	4	3kk	3	3	2	3	3	3	3	3
47	5	3kk	3	4	2	3	3	3	3	2
48	6	3kk	3	3	3	3	3	3	3	2
49	7	3kk	2	2	2	3	3	3	3	2
50	8	3kk	4	3	2	2	2	3	3	2
51	9	3kk	3	3	2	3	2	3	3	3
52	10	3kk	2	3	2	3	3	3	3	2
53	11	3kk	3	2	2	3	3	3	3	2
54	12	3kk	4	3	1	1	3	2	3	2
55	13	3kk	2	3	2	3	3	2	3	1



# Datan tarkistus

- Tärkeä ja aikaavievä vaihe
- Loogisuus – aikuinen ei ole 55 cm pituinen
- Korjaukset täytyy dokumentoida:
  - 18.3.2016 poistettu virheellinen tieto BMI=635.4 (AO)
  - `If ID=890 then BMI=.`;
- Dokumentointi:
  - Muuttujien keräys / mittaus
  - kyselysarjojen tiedot
  - vastausten koodaus





# Datan käyttö

- Alkuperäinen data ‘lukitaan’
- Henkilötiedot erilliseen tiedostoon
- Tutkijoilla käytössä erillinen työkopio aineistosta, jossa ei tunnistetietoja
  - Kyseessä on edelleen kuitenkin henkilötietoja sisältävä pseudonyymidata, joten tietosuojasta pidettävä erityinen huoli
- Kirjanpito luovutetuista datoista ja niiden versioista



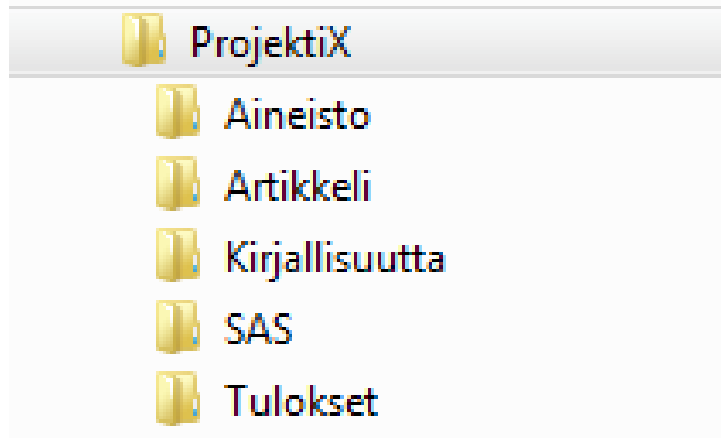
# Tiedostojen hallinta

- Mikä on alkuperäinen ja mikä viimeinen versio aineistosta? Mitä muutoksia siihen tehtiin?
- Entä kun samannimisiä versioita on eri koneilla ja muistitikuilla monta?
- Varmuuskopioitu serveri hajosi, eikä palveluntarjoaja osannutkaan palauttaa tiedostoja.
- SAS-syntakseja on monia, mitä niistä käytettiin artikkelin tulosten tuottamiseen?
- Ohjelmisto on siirtynyt uuteen versioon, eikä vanha tiedosto enää aukea



# Tiedostojen hallinta

- Looginen kansiorakenne
  - Esim. Jokaiselle projektille oma kansio, jonka sisällä tarpeen vaatiessa alakansioita





# Tiedostojen hallinta

- Varmuuskopiointi
  - Verkkolevy jossa varmuuskopiointi
  - Väh. yksi lisäkopio
- Yksi versio jokaisesta tiedostotyyppistä muotoon, joka ei ole riippuvainen ohjelmaversioista.
  - xlsx, sav ja sas7bdat → csv tai txt
  - docx, sas, r → txt, rtf



# Tiedostojen hallinta

- Havainnolliset nimet
  - Alkup\_aineisto.xlsx
  - Jalanpituus.sas
- Aina kun tiedostoa muokataan, muutokset tallennetaan uuteen versioon:
  - Diabetes\_BMJ\_V3.docx



# Käytännön neuvoja

- Avoin tiede ja tutkimus –osaajakoulutus:  
Case: Kokemuksia bioalalta –datan rakenne,  
dokumentointi, management plan ja tarkistukset  
Eliisa Löyttyniemi, biostatistikko  
Kliininen laitos, Turun yliopisto  
<http://avointiede.fi/menneet-osaajakoulutukset> tai  
<http://portti.avointiede.fi/tutkimusdata/hyva-aineistonhallinta-lahtee-ruohonjuuritasolta>



# Tutkimuksen loputtua

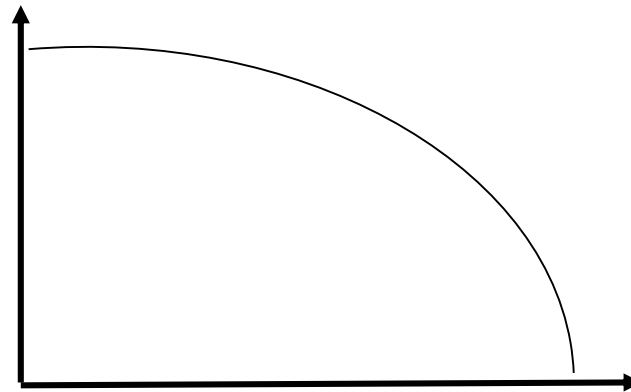
- Arkistointia ja jatkokäyttöä varten tarvittavat tiedostot talteen – muistutus kaikille työkopion saaneille aineiston tuhoamisesta
- Tuhottavaksi vain aineisto, joka on luvattu tuhota tutkimuksen päätteeksi tai josta arvellaan, ettei ole hyötyä myöhemmissä tutkimuksissa



# Tutkimuksen loputtua

- Arkistointia ja jatkokäyttöä varten aineiston tiedot on jo tallessa
- Anonymisoinnissa

Tieteellinen käyttöarvo



Datan turvallisuus



# Kiitos!

*Kysymyksiä, kommentteja?*



TIETOARKISTO  
FINNISH SOCIAL SCIENCE  
DATA ARCHIVE