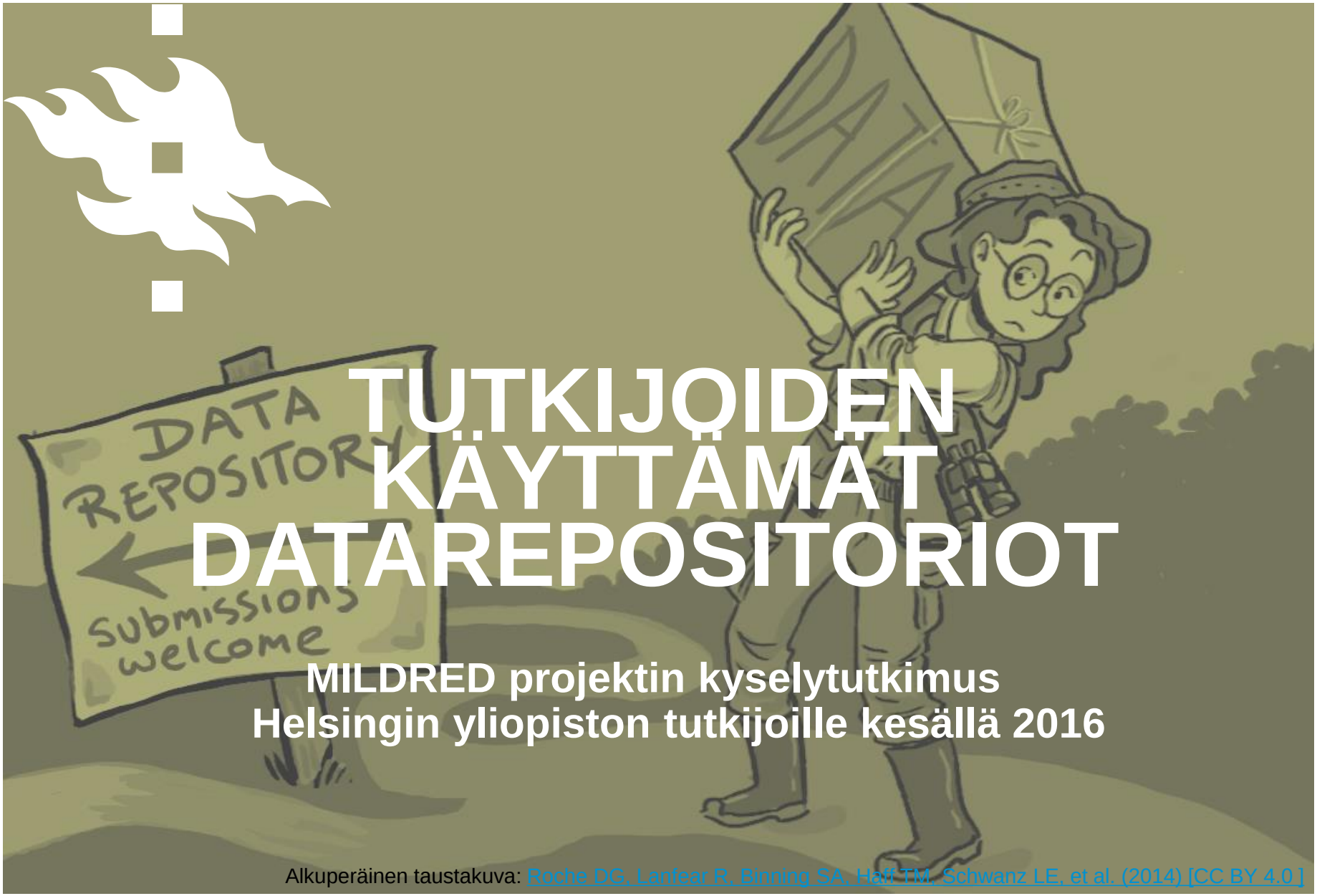


TUTKIJOIDEN KÄYTTÄMÄT DATAREPOSITORIOT

MILDRED projektin kyselytutkimus Helsingin yliopiston tutkijoille kesällä 2016

Alkuperäinen taustakuva: Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) [CC BY 4.0]

ESISELVITYS TUTKIJOIDEN KÄYTTÄMISTÄ DATATIETOKANNOISTA

TOTEUTUS

- Selvitys tehtiin tutkimusdatainfrastruktuurin kehittämishanke MILDRED:n, osaprojektin 3: Julkaisu- ja metadatapalvelut tarpeisiin
- Selvityksen toteutti harjoittelija Anna Salmi informaatikko Mari Elisa Kuusniemen ja kirjastonhoitaja Mikko Ojasen ohjauksessa
- Selvitys sisälsi kolme vaihetta

VAIHE 1: DATAINVENTAARIO

- Kesäkuussa 2016 inventoitiin 250 avoimesti verkossa saatavilla olevan, 2015–2016 ilmestyneen tutkimusartikkelin data.
- Pääotanta PLOS-julkaisusta, mukana myös Nature-julkaisu.
- PLOS-julkaisu edellyttää kirjoittajilta selvityksen tutkimusdatan sijoittamisesta ja avoimuudesta.
- Tuloksia tarkasteltaessa on hyvä huomata, että PLOSin kohdalla Figshare-datatietokanta on erityisasemassa, sillä kustantaja suosittelee sitä julkaisun dataliitteiden pilvijulkaisualustaksi.

VAIHE 2: TUTKIMUSDATAKYSELY

- Kysely suunnattiin sähköpostitse koko Helsingin yliopiston tutkimushenkilökunnalle kesäkuussa 2016
- 258 vastausta
- 62 % elämätieteistä, 21 % humanistisista ja yhteiskuntatieteistä sekä 17 % luonnontieteistä
- Kysely sisälsi monivalintataulukot eri tietokannoista ja vaihtoehtoisista datan säilytyspaikoista ja -laitteista sekä vapaan vastauskentän omia perusteluja varten.

MONIVALINNAN DATAREPOSITORIOT

- ArrayExpress
- Dryad
- European Nucleotide Archive
- Figshare
- FIN-CLARIN
- Finnish Social Science Data Archive
- dbGaP
- Gene Expression Omnibus
- GenBank
- GitHub
- Sequence Read Archive
- UK Data Archive
- Worldwide Protein Data Bank
- Other

MUUN TALLENNUKSEN VAIHTOEHDOT

- University of Helsinki network hard drive
- CSC’s platforms
- Commercial cloud services (Dropbox, SugarSync etc)
- Personal computer hard drive
- External hard drives
- USB memory device
- Other repositories or devices

KYSELYN TULOKSIA

- 44 % vastaajista käytti yhtä tai useampaa tietokantaa
- 21 % käytti kahta tai useampaa, 10 % kolmea tai useampaa
- 56 % ei käyttänyt mitään tietokantaa
- 15 % käytti jotain muuta kuin monivalinnan tietokantoja
- Yleisimpiä olivat GenBank (16,7 %), GitHub (14 %), Sequence Read Archive (6,6 %) ja Gene Expression Omnibus (5 %)

MIKSEI REPOSITORIOITA KÄYTETTY?

- 29 % vastaajista ei tiennyt datan sijoittamismahdollisuuksista tarpeeksi
- 11 %: data oli luonteeltaan sensitiivistä
- 54 % säilytti dataa HY:n verkkolevyllä
- 68 % käytti henkilökohtaista tallennustilaa
- 58 % käytti ulkoisia kovalevyjä
- 50 % käytti USB-tikkua
- 37 % piti dataa kaupallisissa pilvipalveluissa

MIKSEI TARVETTA DATAN SIJOITTAMISEEN?

- 11 %: kysymys epärelevantti omien aineistojen ja tieteenalan kannalta
- 8,5 %: ei tarvetta tarkemmin määrittelemättömästä syystä
- 7,7 %: dataa syntyi vain vähän
- 4,6 %: nykyiset tallennustilat ja -palvelut olivat riittäviä.

POIMINTOJA VASTAUKSISTA

- ” I do not know or trust them [repositories] enough. I do not have such big data that it would be a problem to store it otherways. I would need a system that is reliable, easy to use and access and permanent solution.”
- “The [research] results are fully covered by the published articles.”
- “Unclear benefits with respect to effort.”
- “It was sufficient until this moment to store the data within University infrastructure, although convenient data sharing between collaborators is still lacking.”


METADATASELVITYS

JATKOSELVITYS METADATASTA

- Esiselvityksen ja kyselyn tuloksena kartoitettiin 48 tietokantaa, joissa on HY:n dataa.
- Näille etsittiin Re3data-tietokantarekisteristä tietokannan tunniste ja sen avulla tietokannan API:n kautta haettiin tietokantakohtaiset tiedot: datatyyppi, avoimuus, ohjelmiston nimi, siteerausohjeistus, laatukontrolli, mahdollinen metadatastandardi jne.
- Kustakin tietokannasta haettiin metadatakentän kuvaus. Sitten tietokantojen hakuominaisuuksia testattiin hakemalla henkilön nimellä (tietokantaan tallentamisesta ilmoittanut tutkija) ja organisaation nimellä (HY).

HENKILÖ- JA ORGANISAATIOTIEDOT

- Henkilön nimellä pystyi hakemaan 21/48 tietokannassa.
- Nimellä ei tullut tuloksia 22/48 tietokannassa.
- 4/48 tietokannoista jäi epäselviksi, sillä metadatan kuvausta ei löytynyt eikä testihaku tuottanut luotettavaa tulosta.
- Organisaation nimellä pystyi hakemaan 9/48 tietokannassa.
- Organisaation nimellä ei voinut hakea 36/48 tietokannassa.
- 3/48 tietokannoista jäi epäselviksi, sillä metadatatietoja ei löytynyt eikä testihaku tuottanut luotettavaa tulosta.

REPOSITORIOT, JOISSA MAHDOLLISTA HAKEA ORGANISAATIOLLA

- Database of Genomes and Phenomes (dbGaP)
- GitHub
- Global Biodiversity Information Facility
- Inspire-HEP
- Kielipankki
- MG-RAST (Metagenomics analysis server)
- Tietoarkisto
- Zenodo

YHTEENVETO

- Monet tutkijat (44 % kyselyyn vastanneista) käyttävät kansallisia ja kansainvälisiä datatietokantoja.
- Toisaalta monilla tutkijoilla ei ole riittävästi tietoa asiasta. Lisää palveluja ja neuvontaa tarvitaan.
- Tiedämme tällä hetkellä 48 tietokantaa, joissa on Helsingin yliopiston tutkijoiden dataa. Voimme kuitenkin erottaa Helsingin yliopiston datan muusta datasta vain 9/48 tapauksessa.

YHTEENVETO (2)

Jatkossa MILDRED-hankkeen painopisteitä datapalvelujen kehittämisessä ovat

1) tallennus- ja säilytyspalvelujen kehittäminen
2) datan rikastamisen mahdollistaminen analyysi- ja visualisointivälineiden avulla
3) datan jakamisen ja julkaisemisen mahdollistavat palvelut.

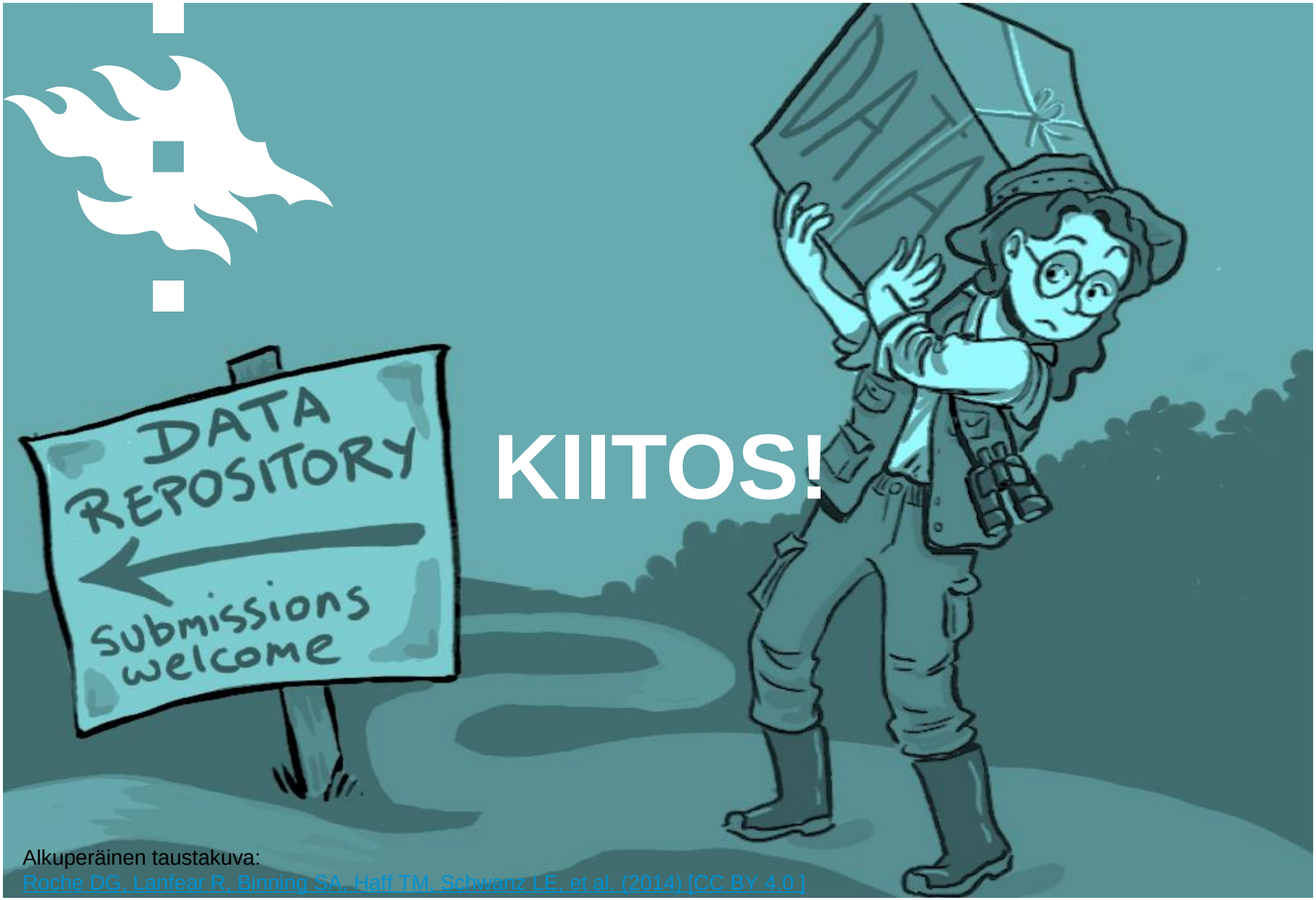
LÄHTEET

Pitkänen, Timo (2016). Storing, sharing and visualizing. Project MILDRED: Development Project of the Research Data Infrastructure at the University of Helsinki [Haettu 27.10.2016]. http://blogs.helsinki.fi/mildred/2016/09/21/storing-sharing-and-visualizing/

Salmi, Anna (2016). Project MILDRED data inventory [unpublished dataset]. University of Helsinki.

Salmi, Anna, Ojanen, Mikko & Kuusniemi, Mari Elisa (2016). Project MILDRED Research Data Repository Survey, University of Helsinki. University of Helsinki. Figshare. DOI: https://figshare.com/articles/Project_MILDRED_Research_Data_Survey/3806394

Salmi, Anna & Pitkänen, Timo (2016). Project MILDRED metadata inventory [unpublished dataset]. University of Helsinki.



KIITOS!

Alkuperäinen taustakuva:
Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) [CC BY 4.0]