Tietojenkäsittelytieteilijöiden kokemuksia

Ella Bingham, tutkimuskoordinaattori,
HIIT, Aalto

Sisältää näkemyksiä

Bioalan / bioinformatiikan tutkijoilta (Datat GB--TB)

Muiden dataintensiivisten alojen tutkijoilta

Aallon IT-henkilökunnalta

Omasta arjesta tutkimuskoordinaattorina ja aiemmin tutkijana

Sisältö

Miten insentivoida tutkijoita

Mitä konkreettisia tarpeita tutkijoilla on

Tallennuspolitiikan esimerkki

Tietokantoja

Miten insentivoida tutkijoita

Tutkijoita pitäisi herätellä data management -suunnitelmien tarpeellisuudesta

Sekä lyhyt- että pitkäaikaissäilytys

Myös pienet datat talteen

Datojen monimuotoisuus huomioitava

Muiden yliopistojen kokemukset olisivat arvokkaita: mitä on tehty, mikä on toiminut?

Seuraavilla kalvoilla ehdotuksia

Palvelun käytöstä on saatava tutkimuksellista kilpailuetua

Keino jakaa omaa dataansa helposti ja meritoitua

Keino järkeistää omaa työskentelyä. Esimerkiksi eri datojen yhteismitallistaminen jo alkuvaiheessa. Työlästä jälkikäteen.

Keino käsitellä, siirtää ja tallentaa dataa paremmin kuin ennen.

Palvelun on toimittava hyvin

Käyttöliittymän on oltava helppo. Sekä komentorivipohjainen että selain- tms. pohjainen.

Laajennettavissa niin että tutkija voi itse rakentaa “kilkkeitä”

Mahdollisimman vähän velvoittavia sääntöjä, pakollisia kenttiä

Best practices koetaan arvokkaiksi

Huomioitava että usein datan omistaa kolmas taho, joka määrää sen tallennuksesta jne.

FIMMin toive: oma tekninen analyysilaite keskustelisi järjestelmän kanssa suoraan

Tarpeita: Datan siirtäminen

Jopa kymmenien TB siirtoa maasta toiseen.

Aallossa ei nyt toimivaa järjestelmää.

Funetin filesender ei ole tarpeeksi järeä ja joustava.

Tarvittaisiin isompi kapasiteetti, järkevä käyttöliittymä ja fiksu protokolla, esim. torrent.

Toiveena myös saada Suomeen kopiot kansainväl. tietopankeista (NIH jne.)

Tarpeita: Lyhytaikaissäilytys

Levytilaa on aina liian vähän.

Mutta pienetkin datat pitäisi saada talteen.

Versionhallinta on parempi kuin vakiovälein tehtävä backup. Hyvä jos automaattinen.

Jopa vuosien ajan käsillä, ei arkistossa.

Aallossa käynnistetty nyt tallennusprosessi.

Tarpeita: Helppo metadata

Kevyt, jotta tulee tehtyä. Jos raskas niin väärinkäytetään tai jätetään käyttämättä

Kattaa minimitiedot: kuka, koska, mikä projekti, julkaisut, kuka saa käyttää ja miten käyttö onnistuu

Mahdollisuus keksiä itse lisää metatietokenttiä

Metatiedot myös datasta joka sijaitsee muualla

Tarpeita: Datan julkaisu

Tutkijalle meritoitumisen keino, että oma data on saatavilla ja siihen viitataan.

Siis tilastoitava, montako kertaa dataa on haettu, ja mistä osoitteesta haku tuli.

Velvoitus että datan hakija viittaa julkaisussaan datan lähteeseen

Viittauksia tulee lisää, jos tutkija kirjoittaa “siteerauskelpoisen” teknisen raportin datasta

Tarpeita: Arkistointi

Useita MB--TB kokoisia aineistoja, joita pitäisi tallettaa 10+ vuotta.

IDA/irods voisi ratkaista

Eettiset ym. syyt joskus rajoittavat pääsyä

Kolmannen osapuolen datoja meillä ei oikeutta arkistoida

Tallennuspolitiikan esimerkki: BECS-laitos

Lääketieteellisen tekniikan ja laskennallisen tieteen laitos, Aalto

Käsittelee suuria määriä eri alojen datoja, erilaisin salassapitorajoituksin

Teki data storing policyn jo 2011

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

Tallennuspolitiikka BECS

Noudatetaan Suomen lakeja ja Aallon käytäntöjä (tässä järjestyksessä).

Tallennuspolitiikka koskee kaikkea laitoksella tallennettua dataa, ellei toisin sovittu. Kolmannen osapuolen datat käsitellään erikseen.

Kaikella datalla on omistaja ja rajallinen elinaika.

Tallennuspolitiikka BECS

Kotihakemisto (Home) on henkilökohtaiselle datalle.

Verkkolevyjen hakemistot (Project, Archive, Scratch) ovat vain tutkimusdatalle. Tiedostot omistaa tutkimusryhmä, ja laitosjohtaja voi saada käyttöoikeudet.

Pöytäkoneiden hakemistoissa on vain tilapäistä dataa ja BECSin ylläpito voi uudelleenasentaa koneet koska tahansa.

Alakohtaisia tietokantoja

National Climatic Data Center, National Oceanic and Atmospheric Administration www.ndcd.noaa.gov

Pangaea, Data publisher for earth & environmental science (Open access) www.pangaea.de

U.S. National Institutes of Health NIH www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

European Bioinformatics Institute www.ebi.ac.uk

GenBank www.ncbi.nlm.nih.gov

Cancer Genomics Hub www.cghub.ucsc.edu

Yleisiä tietokantoja

HY:n kirjaston Opas tutkimusaineistojen tiedonhallintaan (Solmu-projekti) wiki.helsinki.fi/display/aineistonhallinta/Etusivu

NorStore, Norwegian Storage Infrastructure www.norstore.no

The Dataverse Network project, Harvard University thedata.org

Australian National Data Service www.ands.org.au

JISC, Iso-Britannian datapalvelukeskus www.jisc.ac.uk

Kiitokset

Anne Sunikka, Mikko Hakala / Aalto IT

Imre Västrik / FIMM

Antti Honkela, Jaakko Peltonen, Pekka Marttinen, Sami Kaski / Aalto ja HY

TTA:n tietoinfrastruktuurityöryhmä

ym.

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY