Uudelleenkäytettävää, oikein kuvattua ja laadukasta dataa

Tutkijayhteisön luomia työkaluja ketterään datanhallintaan

Huolellinen datanhallinta mahdollistaa laadukkaan tutkimuksen nyt ja tulevaisuudessa. Sen hallinnalle on luotu ns. FAIR-periaatteet eli data on löydettävissä, se on saatavissa, se on yhdistettävissä muihin vastaaviin datakokonaisuuksiin ja sitä voidaan käyttää uudestaan. Näiden periaatteiden pohjalta ELIXIR-infrastruktuuri tarjoaa käyttöön hyviä datanhallintatyökaluja, jotka tukevat tutkijaa datanhallinnan eri vaiheissa.

“Hyvään tieteelliseen käytäntöön kuuluu varmistaa datan säilyminen käyttökuntoisena ja hyvin dokumentoituna koko tutkimusprosessin ajan ja siten, että tutkimustulokset ovat todennettavissa tutkimusprosessin jälkeenkin. On tärkeää, että tutkijat ja tietojärjestelmät voivat löytää ja saada käyttöönsä yhteen toimivia ja uudelleenkäytettäviä tutkimusaineistoja. Tämän edistämiseksi vuonna 2016 julkaistiin FAIR-periaatteet”, sanoo CSC:n datanhallinnan asiantuntija **Minna Ahokas**.

”ELIXIRin tarjoamien ohjeiden ja työkalujen avulla tutkijan on helpompi tehdä datansa löydettäväksi, saavutettavaksi, yhteentoimivaksi ja uudelleenkäytettäväksi ja samalla noudattaa datanhallinnassaan FAIR-periaatteita.”

Yhteistyössä eri jäsenmaiden ELIXIR-keskusten kanssa luotu RDMkit-sivusto pyrkii tukemaan ja yhtenäistämään datanhallinnan käytäntöjä Euroopassa.

RDMkit sisältää ohjeita ja vinkkejä aineiston koko elinkaaren ajalle: datanhallinnan suunnittelusta ja data-analyyseistä aina datan julkaisemiseen ja uudelleenkäyttöön.

”RDMkit on toteutettu niin, että jokainen joka on tekemisissä datan kanssa voi ottaa sen työkalupakikseen. Se tarjoaa ohjeistuksen lisäksi linkit palveluihin, joita tutkija tai tutkimuksen tukipalveluissa työskentelevä tarvitsee datanhallinnan eri vaiheissa”.

Suomen ELIXIR-keskus eli CSC on yksi niistä, joka tuottaa sisältöä ja ylläpitää työkalupakettia.

Ahokas korostaa, että sivustoa on suunniteltu alusta lähtien läpinäkyvästi yhteistyössä tutkijoiden ja datanhallinnan asiantuntijoiden kanssa. Kuka tahansa ELIXIR-infrastruktuuriin kuuluvista voi osallistua kehitystyöhön. Kaikki on dokumentoitu ohjelmakehitys-projekteille tarkoitettuun GitHub-portaaliin.

”RDMkitissä dataa tarkastellaan sen elinkaaren vaiheiden kautta. Datan keräämiseen, kuvailuun tai julkaisemiseen on tarjolla omat ohjeensa.”

RDMkit kehitettiin ELIXIR-CONVERGE -hankkeessa. Datanhallinnan yhtenäistämiseen oli tarvetta, koska tutkimushankkeet ovat pääasiassa kansainvälisiä ja dataa liikutellaan kansallisten rajojen yli.

”RDMKit on ensimmäinen iso kansainvälinen yritys yhtenäistää datanhallinnan

käytäntöjä ja ohjeistuksia, jotta saadaan uudelleenkäytettävää, sekä riittävästi, yhtenäisillä standardeilla kuvailtua ja laadukasta dataa. Datanhallinnassa on kyse siitä, että datan keruu, käsittely ja kuvailu suunnitellaan ajoissa: miten ja missä dataa säilytetään ja miten eri versioita hallitaan. Sitten on vielä mietittävä, onko datassa jotakin sellaista, mikä pitäisi säilyttää pitkäaikaisesti. Toisaalta pitäisi myös päättää, mikä osa datasta voidaan hävittää."

Minna Ahokkaan mielestä on tärkeää tarjota tutkijoille palveluja, jotka auttavat heitä noudattamaan datanhallinnan hyviä käytäntöjä.

"Yritämme välttää tilannetta, että tutkijoille esitetään esimerkiksi rahoitushakujen yhteydessä aina uusia listoja datanhallinnan vaatimuksista, mutta ei osoiteta niihin sopivia palveluita. Jos vaadimme, että tutkimushankkeiden datanhallinnassa noudatetaan FAIR-periaatteita, meidän pitää tarjota riittävästi tukea ja palveluita FAIR-datan tuottamiseen."

Asiantuntijatukea datanhallintaan

CSC, suomalaiset tutkimusorganisaatiot ja yliopistot ovat luoneet kansallisen datatuki-verkoston. Verkosto toimii CSC:n ja organisaatioiden datatukihenkilöstön yhteistyön tukena. Se tarjoaa foorumin avoimelle keskustelulle, kysymysten esittämiselle ja vertaistuelle.

Esimerkiksi Aalto-yliopistossa on lanseerattu tieteenalakohtaiset "data-agentit", jotka ovat datanhallinnan asiantuntijoita ja heillä on tutkijatausta. He huolehtivat yhdessä tutkijoiden kanssa datasta.

RDMkitin julkaisuvaiheessa datanhallintaan kohdistui COVID 19 -pandemian vuoksi aivan uudenlaisia paineita.

"Kun RDMkit oli saatu lähes valmiiksi, maailmaan iski COVID. Silloin totesimme ELIXIR-CONVERGE -hankkeessa, että myös COVID-virukseen liittyvä data ja sen vaatimukset pitää huomioida. Siksi RDMkitiin toteutettiin nimenomaan COVID 19 -datan käsittelyyn liittyvää ohjeistusta sekä Euroopan COVID 19 -dataportaalia koskeva sivu."

RDMkit ja ELIXIRin datanhallinnan ohjeistukset ovat päätyneet myös osaksi EU:n Horizon Europe rahoitusinstrumentin datanhallinnan ohjeita.

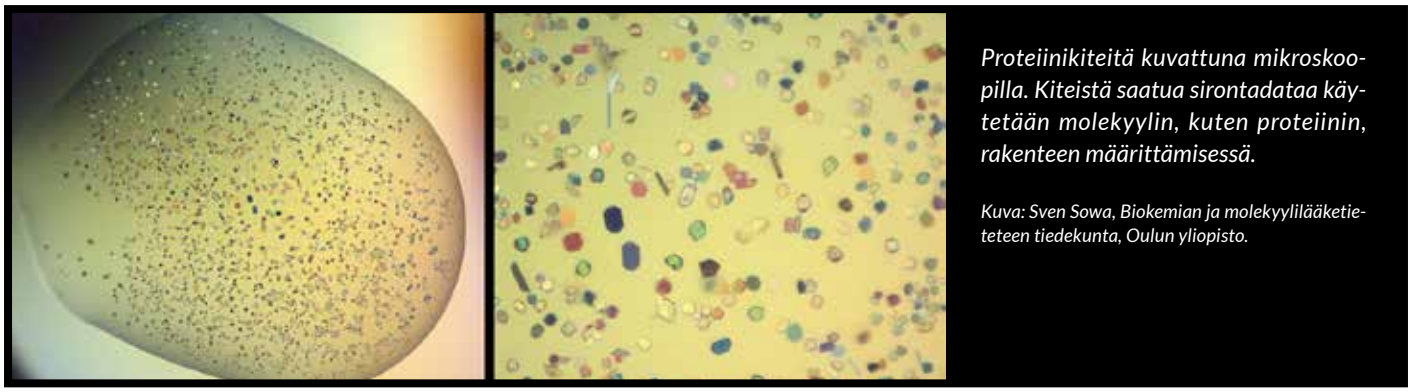
Biotietieteissä suositellaan RDMKit-työkalupakin hyödyntämistä. Se on herättänyt myös maailmanlaajuista kiinnostusta. Yhdysvaltalaisia käyttäjiä on huomattava määrä ja NIH (National Institutes of Health) on kiinnostunut yhteistyöstä ELIXIR-infrastruktuurin kanssa.

IceBear – sovellus rakennebiologian datanhallintaan

RDMkit on yleinen datanhallinnan ohjekokoelma, josta linkataan eteenpäin esimerkiksi IceBearin kaltaisiin työkaluihin.

"IceBear on tehty alun perin kristallografiaa ja rakennebiologian tiedonhallintaa varten", sanoo rakennebiologian professori **Lari Lehtiö** Oulun yliopiston biokemian ja molekyylilääketieteen tiedekunnasta.

Lehtiö on myös rakennebiologian tutkimusinfrastruktuurin Instructin Oulun yksi-



Proteiinikiteitä kuvattuna mikroskoopilla. Kiteistä saatua sirontadataa käytetään molekyylin, kuten proteiinin, rakenteen määrittämisessä.

Kuva: Sven Sowa, Biokemian ja molekyylilääketieteen tiedekunta, Oulun yliopisto.

kön johtaja. Biocenter Oulun rakennebiologian yksikössä suunniteltiin professori **Rik Wierengan** ja sovelluskehittäjä **Ed Danielin** avulla rakennebiologian datanhallintaohjelma IceBear. Sovelluksen kehitystyötä on tehty myös ELIXIRin koordinoimassa EOSC-Life verkostossa, johon myös Instruct kuuluu. EOSC-Life projektin tuella IceBear siirrettiin CSC:n ylläpitämään cPouta-pilvipalveluun.

Biocenter Oulussa kiteytetään proteiineja ja muita makromolekyylejä. Proteiinien aminohappoketju on laskostunut kolmiulotteiseksi rakenteeksi, joka on kullekin proteiinille tyypillinen. Koska mahdollisia erilaisia laskostumisen tapoja on valtavasti, proteiinirakenteita on jouduttu selvittämään laboratorioissa kokeellisesti, kiteyttämällä. Proteiinin kolmiulotteinen rakenne pystytään selvittämään sen perusteella miten röntgensäde siroaa proteiinikiteestä. Kerätystä sirontadatasta voidaan matemaattisella muunnoksella laskea proteiinin elektronitiheyskartta, joka kertoo atomien paikat proteiinissa. Nykyään käytetään rakennetutkimuksessa paljon myös kryoelektronimikroskopiaa, jossa proteiineista valmistettua jäädytettyä näytettä pommitetaan elektroneilla ja miljoonat yksittäiset proteiinien 2D-kuvat yhdistetään kolmiulotteiseksi rakenteeksi.

Apuna proteiinien kiteytyksessä on automaattisia kuvantamislaitteistoja. Proteiinit kiteytetään eri liuoksissa, jolloin joissakin olosuhteissa tapahtuu kiteiden muodostuminen.

”Proteiini kiteytetään pisaraan ja tätä seurataan kuvantamalla. Levyissä voi olla 300 pisaraa ja levyjä monta sataa. Kun niitä kuvataan joka päivä, kuvia tulee aika paljon. Kiteytys tehdään yleensä roboteilla,” sanoo Lehtiö.

Kidenäytteet poimitaan käsin mikroskoopista ja laitetaan nestetyppitankkeihin. Nyt IceBear-ohjelman avulla voidaan samalla sujuvasti pitää kirjaa automaattisesti näytteistä ja niihin liittyvästä tiedosta.

”Usein näytteet lähetetään toiseen infrastuktuuriin, eri synkrotroneihin Eurooppaan. IceBearin avulla tiedetään, mitä näytteelle tapahtui toisessa paikassa. Metadataa liikutellaan eurooppalaisten synkrotronien käyttämien tietokantojen ja IceBearin välillä. Näytteessä on metadataa aika paljon, kuten mikä proteiini oli kyseessä ja millainen rakenne sillä oli, miten se kiteytettiin ja minkälaiset olosuhteet kiteytyksessä olivat.”

Icebearin avulla päästään eroon käsin tehdystä kirjanpidosta. Dataa voidaan lähettää ilman kaavakkeiden täyttöä ja linkit ovat tietoturvallisesti luotu näytteiden viivakoodeihin.

“Kun tämän tekee kerran, se on siinä. Tämän sovelluksen arvo esimerkiksi tutkijoiden ajan säästämisessä näkyy myös vuosien kuluttua”, sanoo Lehtiö.

Ari Turunen

LISÄTIETOJA:

RDMkit
https://rdmkit.elixir-europe.org/covid19_data_portal

ELIXIR CONVERGE
https://elixir-europe.org/about-us/how-funded/eu-projects/converge

COVID-19 dataportaali
https://www.covid19dataportal.org

EOSC-Life
https://elixir-europe.org/news/eosc-life-start

IcebBear
https://icebear.fi/

CSC – Tieteen tietotekniikan keskus Oy

on valtion omistama, opetus- ja kulttuuriministeriön hallinnoima, voittoa tavoittelematon osakeyhtiö. CSC ylläpitää ja kehittää valtion omistamaa keskitettyä tietotekniikkainfrastruktuuria.

http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR

rakentaa infrastruktuurin bioalan tutkimuksen tueksi. Se yhdistää 21 Euroopan maan ja Euroopan molekyylibiologian laboratorion EMBL:n johtavat organisaatiot yhteiseksi biologisen informaation infrastruktuuriksi. Sen Suomen keskus on CSC Tieteen tietotekniikan keskus Oy.

http://www.elixir-finland.org
http://www.elixir-europe.org

SUOMEN ELIXIR
Puh. +358 9 457 2821 e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR PÄÄMAJA
EMBL-European Bioinformatics Institute
www.elixir-europe.org