Datanhallinnan ja laskennan kehittämisohjelma

minedu.fi/DL2021

Datanhallinnan ja laskennan tutkimusinfrastruktuurien, palveluiden ja osaamisen kehittämisohjelma 2017–2021

Yhteenveto tutkijoiden ja tutkimusorganisaatioiden työpajasta 13.11.2017

Johdanto	2
Ryhmä A (esim. bioinformatiikka, life science, tietojenkäsittelytiede)	3
Ryhmä B (esim. fysiikka, kemia, nanotieteet, geotieteet)	4
Ryhmä C (esim. tutkimuslaitokset, ammattikorkeakoulut)	4
Ryhmä D (esim. digitaaliset ihmistieteet)	5
Ryhmä E (esim. tietojenkäsittelytieteet, tutkimuksen it-tuki)	6
Liite 1. Ohjelma	8
Liite 2. Pienryhmissä käsiteltävät teemat	9

Lisätietoja:

Anu Nuutinen
DL2021-ohjausryhmän asiantuntijasihteeri
Johtava tiedeasiantuntija, Suomen Akatemia
etunimi.sukunimi@aka.fi

Datanhallinnan ja laskennan kehittämisohjelma

minedu.fi/DL2021

Johdanto

Opetus- ja kulttuuriministeriö toteuttaa tutkimus- ja innovaatiotoimijoiden kanssa datanhallinnan ja laskennan tutkimusinfrastruktuurien, palveluiden ja osaamisen kehittämisohjelman vuosina 2017−2021. Kehittämisohjelmassa investoidaan noin 35 miljoonaa euroa datanhallinnan ja laskennan infrastruktuureihin sekä niihin liittyviin palveluihin.

Kehittämisohjelmalla päivitetään CSC – Tieteen tietotekniikan keskus Oy:n infrastruktuureja kansainvälisen yhteistyön varmistavalle tasolle. Uudistettava infrastruktuuri ja siihen liittyvät palvelut tarjotaan tutkimusyhteisön käyttöön aiempaa laajemmin. Kehittämisohjelmassa huomioidaan erityisesti alan eurooppalainen tutkimusinfrastruktuuripolitiikkakehitys.

DL2021-kehittämisohjelman 13.11.2017 järjestämässä työpajassa yli 50 tutkijaa ja tutkimushallinnon asiantuntijaa yliopistoista, ammattikorkeakouluista sekä valtion tutkimuslaitoksista työstivät tilannekuvaa datanhallinnan ja laskennan tutkimusinfrastruktuuriin, palveluihin ja osaamiseen kohdistuvista tarpeista (tilaisuuden ohjelma löytyy *liitteestä 1*).

Työskentely toteutettiin viidessä pienryhmässä, jotka tarkastelivat datanhallinnan ja laskennan tutkimusinfrastruktuuria, palveluja, osaamista, yhteistyötä ja priorisoinnin periaatteita edustamillaan aloilla ja tutkimusorganisaatioissa (teemoihin liittyvät tarkemmat kysymykset on kuvattu *liitteessä 2*). Tässä koosteessa esitetään tiivis yhteenveto kunkin ryhmän keskustelusta.

Ryhmä A (esim. bioinformatiikka, life science, tietojenkäsittelytiede)

- On tärkeää löytää balanssi datan tuottamisen, jakamisen ja säilöntään liittyvien palvelujen välille: data, suurteholaskenta, GPU-laskenta ym. uudet teknologiat.
- Osaamispuute esim. GPU-laskennassa yliopistoissa. CSC:ltä tarvitaan asiantuntija-apua ja tukipalveluja, jotta jatkossa yliopistojen IT:t voivat toimia tukena.
- Käyttäjäkoulutusta tarvitaan esim. pilvipalveluissa.
- Sensitiivisen datan säilöntään ja jakamiseen liittyvät palvelut.
- Tulevaisuudessa ns. data-streaming -palveluita esim. SOME tietojen poimiminen tietovirrasta, kun laskenta tulee enenevässä määrin SSH-aloille, esim. Suomi24-tietojen poimiminen. Mikä on yksityistä dataa ja miten sitä voidaan yhdistellä?
- CSC:lla on rooli keskitettynä järjestelmäkehittäjänä ja kouluttajana, tutkimusorganisaatiot tuottavat tutkimuksen IT:n lähipalvelut ja ylläpitävät osaavaa henkilökuntaa.
- Tutkimuksen IT-tuen resursointi strategisemmaksi ja kestävämmäksi organisaatioissa.
- Yhteistyötä edistävät alustat olisivat tärkeitä! Voisivat olla kilpailuetu Suomelle. Suomella on ainutlaatuinen mahdollisuus, jos pystytään rakentamaan alustoja, joissa voi yhdistää dataa eri lähteistä. Mahdollistaa esim. yritysyhteistyötä (esim. my data).
- Huipputieteellä pitäisi olla resurssit kunnossa laskennan ja muun datapalvelun suhteen.
- Tukea sekä aloitteleville että kokeneille käyttäjille (lisää resursseja saa, kun voi osoittaa tutkimuksen merkittävyyden).
- Datanhallintasuunnitelmia ja niiden laadintaa tukevia prosesseja tulee kehittää.
- CSC:n pitäisi entistä paremmin pystyä kertomaan saavutuksista, joita sen tarjoamat palvelut ovat mahdollistaneet. Laskentaresursseja hakevia voisi velvoittaa antamaan näitä tietoja.

Ryhmä B (esim. fysiikka, kemia, nanotieteet, geotieteet)

- Supercomputers a must: Crucial for enabling high-level science!
  - Provide a competitive edge (also attracts top talent to Finland); “CSC biggest research infrastructure in Finland”.
  - Needed for international highest-level collaborations, enables H2020 participation.
- Distributed (local) clusters crucial addition
  - High-throughput mid-level calculations cost-effectively
  - Can be used as Grid/Cloud/locally
  - Can support local high-memory needs
- Data handling a challenge for medium-term and long-term
  - Currently no good solution for PB-level long-term storage
  - User activation a challenge: services should be made easier to use than now
- GPU and other specialized solutions important for A.I.
- Support
  - CSC training and education activities excellent
  - Online courses/webinars could be increased – even MOOC’s ?
  - CSC voisi koordinoida koulutusta, yhdistää eri yliopistojen osaamisen ja huolehtia tietotaidon levittämisestä.
- Collaboration
  - PRACE important, some existing networks (CERN, Nordugrid etc.) work very well
  - For data handling, what should be the role of national vs. EU-level solutions?
- Resource division by CSC
  - Normal computing project division excellent, lightweight is good
  - Transparency of CSC Grand challenge selection could be increased

Ryhmä C (esim. tutkimuslaitokset, ammattikorkeakoulut)

- Tutkimusinfrastruktuurin monipuolisuus ja joustavuus tärkeää. Nykyiset laskentapalvelut on koettu joustaviksi, tärkeää että datanhallinnan palvelut ym. myös joustavia.
- Datamäärien kasvu, kuvat ja videot.
- Nykytilan mukaan ei voi tehdä päätöksiä, tulee nähdä yli 5 v. aikajänteen.
- Uudet laitteet kuten dronet, satellittikuvien lisäksi myös muut kuvat.
- Pitkät aikasarjat arvokkaita. Erilaisten datojen yhdistäminen tulee yleistymään.
- Data lisää myös laskentaintensiivisyyttä.
- Tuotettujen datojen liikuttaminen on turhaa, tuottajan olisi hyvä vastata datan ylläpidosta. Vanhojen datojen ylläpidon ongelma.
- AMK:ssa syntyy eri tyyppisiä datoja, ei määrällisesti kuitenkaan paljon. Ympäristödataa eniten. Paljon on myös laadullista aineistoa, big data -tyyppistä. Kuvia tulee paljon mutta on myös videoaineistoa. Datan avoimuuden ja saatavuuden turvaaminen esim. 10 vuoden päähän on hankalaa.

- Mahdollisuuksien hahmottaminen hankalaa. Mihin kannattaisi oman laitoksen ja organisaation osaamista kehittää?
- Tutkimuslaitoksilla kautta linjan haasteina budjettirahoituksen pienentyminen ja Valtori. Tärkeää huomata, että hankemaailmassa tutkimusinfrastruktuuria ei saada koskaan kehitettyä.
- Yritysyhteistyö tulee ottaa huomioon. Samaa tutkimusta tehdään molemmin puolin aitaa. CSC:n osalta lisenssiongelmia (kaupallinen/business lisenssit).
- Softien lisenssien tilalle open source softat.
- Miten kansainväliset yhteistyökumppanit pääsevät mukaan infrastruktuurin käyttäjiksi?
- CSC:n kautta pääsy kansainvälisiin resursseihin on tärkeää. PRACE on tärkeä ja sen säilymien tulee turvata.
- Yleiset työkalut CSC:llä, erityistarpeet kullakin sektorilla erikseen. CSC voisi hoitaa perustoiminnon ja erikoisosaaminen tulee laitoksilta. Mutta rahoitusta erikoisinfraan ei löydy.
- Jokaisen tieteenalan ei tarvitse saada omaa datanhallinnan asiantuntijaa CSC:lle.
- Opetustilanteet vaativat eri tavalla läsnäolevaa käyttötukea.

Ryhmä D (esim. digitaaliset ihmistieteet)

- Sensitiiviset aineistot ja data.
- Kapasiteettikysymykset ja avoimuuteen liittyvät peruskysymykset. Erityisiä haasteita: videomateriaalit, paikkatiedot, some-aineistot jne. Säilyttäminen ja anonymisointi. Avoimuuden edistäminen myös julkisen ja yksityisen sektorin yhteistyössä. Yrityssalaisuudet, potilastiedot jne.
- Isot rekisteriaineistot ja laskentakapasiteetin turvaaminen. Pilvipalvelujen kapasiteetti ei riitä. Kehittäminen mutkikasta. Laskentakapasiteetin tarve kasvaa. Myös ison datan hyödyntämiseen liittyvän osaamisen turvaaminen. Etäkäyttö.
- Suomalaisen olemassa olevan kulttuuripääoman tehokas hyödyntäminen ja tunnettuuden edistäminen. Pitkäaikaissäilytyksen kysymykset.
- Miten konkreettisesti kuvataan käyttöoikeudet, -ehdot ja -valtuudet?
- Koulutus. Perusvalmiudet erilaisten aineistojen käyttämiseen. Yleiset tietoturvaan liittyvät kysymykset, sopimustekniset kysymykset, kryptaaminen jne.
- CSC:n tarjoamien palvelujen näkyvyyden ja tunnettavuuden parantaminen yliopistoissa, korkeakouluissa ja muissa tutkimusorganisaatioissa.
- Käyttäjätuki. Erilaiset tarpeet ja valmiudet eri organisaatiossa. Miten saadaan tieto jalkautumaan sinne missä sitä tarvitaan? Toimintakulttuurin kehittäminen.
- Linkittyy saumattomasti alan koulutustarpeisiin. Uusi osaaminen ja uudet taidot huomioitava paremmin tutkijankoulutuksessa.
- Uusi tietoturvaan ja direktiiveihin liittyvä osaaminen ja asiantuntemus. Ymmärrys siitä, mitä datan hyödyntäminen käytännössä tarkoittaa. Sopimukset, metadatan luominen jne.

- Kansainvälinen yhteistyö. Miten VALTTI, VALTORI ja SOTE-aineistot toimivat kansainvälisessä yhteistyössä? Myös kotimaassa eri sektorien välisessä yhteistyössä. Näiden osalta julkinen sektori on OK, mutta miten yksityinen yrityspuoli? Vastauksena käytettävyyteen liittyvät sopimukset.
- Olemassa olevien aiheeseen liittyvien ohjelmien, alustojen ja verkostojen hyödyntäminen yhteistyössä ja tunnettavuuden edistäminen.
- Kansalaisten, yritysten, järjestöjen ja muiden eri toimijoiden avoin osallistaminen. Kansalaistieteen tarjoamat mahdollisuudet.

Ryhmä E (esim. tietojenkäsittelytieteet, tutkimuksen it-tuki)

- Huomioitavaa: ketteryys hankintapuolella, uudistumisen tarve, roolit.
- Haaste: fragmentoituminen (omat viritykset ym.) - miten voidaan yhdessä hallinnoida?
- Tutkimusorganisaatioissa erilaisia malleja tukipalvelujen organisointiin, sekä keskitystä että hajautusta, yhteispeli tärkeää sekä tutkimusorganisaatioissa että eri tutkimusorganisaatioiden välillä ja suhteessa CSC:hen ym. tukipalveluihin.
- Infrastruktuuri vanhenee nopeasti (GPU-arkkitehtuuri muuttuu nopeasti), miten huolehditaan ajantasaisuudesta.
- Uudentyyppiset käyttöliittymät ja näyttöliittymät.
- Datan kuvailuun ja ymmärtämiseen liittyvien palvelujen tarve tunnistettu (esim. kasvatustieteilijät ja kielitieteilijät).
- Reunalaskenta (reaaliaikainen laskenta) tulossa (5G, 6G): Mikä on CSC:n puolella ja mikä on tutkimusorganisaatioiden vastuulla?
- HPC-palveluissa jatkuvuus tärkeää, ennakoitavuus. Toisaalta uudentyyppiset palvelutarpeet nousevat rinnalle.
- Tutkijakokemus erityisesti ”uusilla” aloilla, käyttö helppoa tutkijalle niin, että hän saa nopeasti tehtyä sen mitä on tarpeen.
- Käyttäjämäärät kasvavat, tallennuskapasiteetin tarve kasvaa -> haasteita lähituen rakentamisessa.
- Bootstrap-vaihe, jossa tarvitaan panostuksia tukeen ym. – koulutus ml. menetelmät ja algorimiosaaminen.
- HPC-laskenta ei katoa, mutta tulossa uudenlainen käyttäjäpopulaatio, jonka tarpeet ovat hyvin erilaisia. Iso data. Pieni data. Uudenlaiset käyttöskenaariot.
- Asiantuntija/yhteistyöverkosto, esimerkkinä FGCI-verkosto, jossa viikoittainen yhteydenpito, ponnahduslauta CSC:n infraan.
- Koulutusta myös maakuntiin, pois pääkaupunkikeskeisyydestä.

- Palvelut 5 vuoden päästä
    - Työkalut oman datan hallintaan, visualisointiin yms. millä tahansa tieteenalalla, helposti ja sujuvasti käytettävissä.
    - Reunalaskennan tuki. Millaiset alustat ja standardit? Kuka hallinnoi/maksaa/missä ajetaan?
    - Koko data-management-ketju kuntoon.
    - Avoimen datan hallinnointi ja käyttö, ml. kansalaistiede.
    - Käyttöliittymä dataan tulee muuttumaan suuresti loppukäyttäjän kannalta.
    - Kokeilut: tehdä jotain mitä muualla ei ole jo tulossa käyttöön.

Liite 1. Ohjelma

Tutkijoiden ja tutkimusorganisaatioiden työpaja

Aika: Maanantai 13.11.2017 klo 12.30–16.00
Paikka: Paasitorni, Paasivuorenkatu 5 A, Hakaniemi, Helsinki (Siltasaari-sali, 1. krs.)

Työpajan tavoitteena on koota mahdollisimman konkreettinen tilannekuva datanhallinnan ja laskennan tutkimusinfrastruktuuriin, palveluihin ja osaamiseen kohdistuvista tarpeista.

Ohjelma

12.30–12.45 Työpajan avaus: Kehittämisohjelman tausta ja tavoitteet
Erja Heikkinen, tiedeasianneuvos, ryhmän päällikkö, opetus- ja kulttuuriministeriö

12.45–13.00 CSC:n kehittyvät palvelut tutkijoille tulevaisuudessa
Tiina Kupila-Rantala, varatoimitusjohtaja, CSC – Tieteen tietotekniikan keskus Oy

13.00–13.05 Pienryhmätyöskentelyn esittely
Anu Nuutinen, johtava tiedeasiantuntija, Suomen Akatemia

13.05–13.10 Siirtyminen pienryhmiin

13.10–14.10 Pienryhmätyöskentely

RYHMÄ A (Petri Myllymäki, pj.)
RYHMÄ B (Kai Nordlund, pj.)
RYHMÄ C (Hannele Korhonen, pj.)
RYHMÄ D (Hannu Salmi, pj.)
RYHMÄ E (Susanna Pirttikangas, pj.)

Jokaisella pienryhmällä on asiantuntijasihteeri, joka vastaa keskustelun raportoinnista.

14.10–14.30 Kahvia tarjolla Paasiravintolassa (2. krs.)

14.30–15.10 Pienryhmätyöskentely jatkuu samoissa kokoustiloissa

15.10–15.15 Siirtyminen pääsaliin

15.15–16.00 Yhteenveto työpajan annista
Pienryhmien puheenjohtajat: Keskeiset nostot jokaisesta ryhmäkeskustelusta (30 min)
Yleisökysymykset (10 min)
Mitä kehittämisohjelmassa tapahtuu seuraavaksi? (5 min)

Liite 2. Pienryhmissä käsiteltävät teemat

Työskentelyn tavoite: Pienryhmää pyydetään tuottamaan mahdollisimman konkreettinen tilannekuva datanhallinnan ja laskennan tutkimusinfrastruktuuriin, palveluihin ja osaamiseen kohdistuvista tarpeista edustamillaan aloilla / tutkimusorganisaatioissa.

- Palvelukokonaisuuden linjaamiseen ja tarpeiden priorisointiin keskeisesti vaikuttavat asiat
- Jokaista pienryhmää pyydetään käsittelemään kaikkia 5 teemaa, mutta ryhmä voi painottaa niiden sisällä keskeisiä näkökulmia kokoonpanonsa mukaan.

1. Datanhallinnan ja laskennan tutkimusinfrastruktuuri

- Dataintensiivisen laskennan erityisvaatimukset infrastruktuureille
- Datan analysointia, hallintaa, ja hyödyntämistä tukevat alustat
- Big data -ratkaisut eri aloilla
- Laskennalliset mallit ja niihin liittyvät laskenta- ja datan säilytyskapasiteetit
- Suurteholaskennan uudet resurssit, ohjelmistojen ylläpito tulevaisuudessa, datan säilyttämiseen liittyvät ratkaisut
- Millaisia datanhallinnan ja laskennan erityiskysymyksiä on uusilla/kehittyvillä tutkimusaloilla?
- Millaisia datanhallinnan ja laskennan erityiskysymyksiä on erilaisilla tutkimusorganisaatioilla?
- Laskentakapasiteetin jakautuminen eri käyttäjäryhmien kesken, jonotusajat
- Hinnoittelu ja kustannustaso
- Toimintavarmuus (käyttökatkot)
- Datanhallinta tutkimuksen elinkaaren eri vaiheissa, avoin data, tietoturva

2. Palvelut

- Käyttäjätuki: mitä CSC:llä, mitä tutkimusorganisaatioissa?
- CSC:n tarjoamat koulutuspalvelut

3. Osaaminen

- Millaista osaamisen vahvistamista kasvava dataintensiivisyys ja kehittyvät tutkimusmenetelmät edellyttävät alallanne / organisaatiossanne?
- Tieteenalakohtaisten dataosaajien tarve (erityisosaajat)
- Substanssiosaajien osaamisen kehittäminen (esim. tutkijankoulutuksessa)

4. Yhteistyö

- Millainen datanhallinnan ja laskennan infrastruktuurikokonaisuus palveluineen mahdollistaa korkeatasoisen kansainvälisen yhteistyön?
- Suomen kiintiöt kansainvälisissä datakeskuksissa ja niiden käyttö
- Eri organisaatioiden yhteistyötä tukevat alustat
- Millainen merkitys yritysyhteistyöllä on datanhallinnan ja laskennan näkökulmista tutkimusalallanne / organisaationne edustamilla tutkimusaloilla?

5. Priorisoinnin periaatteet

Mitä näkökulmia ja miten ohjausryhmän tulisi painottaa hankintaa linjatessaan?

- Tutkimuksen monimuotoisuus: Laskenta ---- Suurteholaskenta
- Laskentapalveluiden/ suurteholaskennan edellytykset datanhallinnalle
- Resurssijaon läpinäkyvyys
- Alan tutkimuksen laatu, uudistuminen ja vaikuttavuus
- Kansainvälisen yhteistyön merkitys alalla: kansallinen infrastruktuuri yhteistyön mahdollistajana ja osaajien houkuttelijana
- Laskentainfrastruktuurin käyttäjäkunnan laajuus alalla (nyt/tulevaisuudessa)
- Datanhallintaan ja laskentaan liittyvä osaamistaso alalla
- Muut näkökulmat?