

Tutkimuksen tietoaaineistot – poikkihallinnollinen haaste

Kansleri Ilkka Niiniluoto



TIETEEN JULKISUUS

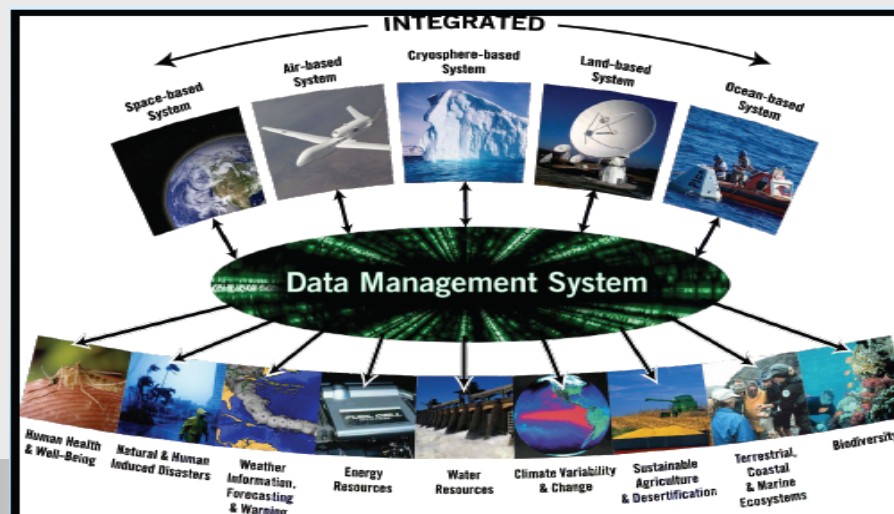
- **tiede on julkisesti perusteltua tietoa**
 - havainnot, data, kokeiden toistettavuus
- **painetut julkaisut, referee-järjestelmä**
- **ICT:n mahdollisuudet**
 - elektroninen julkaiseminen
 - tutkimustiedon avoin saatavuus (OA), digitaaliset julkaisuarkistot
 - tutkimusaineistojen tallennus
 - e-science, tieteellinen laskenta, simulointi
- **yliopistot, tutkimuslaitokset, tieteelliset seurat, kirjastot, arkistot**

Tutkimusaineistojen kasvava rooli

- Tieteellinen tutkimus tuottaa ja käyttää yhä enemmän sähköisessä muodossa olevaa tai siihen muunnettua primaariaineistoa
- Näitä ovat mm. havainto- ja mittaustiedot, mallinnus- ja analyysitulokset sekä rekisteri- ja tilastotiedot

Visio: kaikenlaisten tieteellisten sisältöressurssien olisi oltava helposti saatavilla käyttäjäystävällisten e-infrastruktuuripalvelujen välityksellä ja hyödynnettävissä tietämyksen jakamista silmällä pitäen.

(Komission vuonna 2009 antama tiedonanto "Tieto- ja viestintätekniikka tieteen palveluksessa" (KOM(2009)108))



Tutkimuksen tulevaisuudessa tuottamia datamääriä

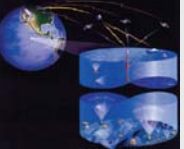
- **LHC: 16 PB/vuosi**
- **IPCC AR5 2010: 10 PB/malli**
- **EOSDIS Satelliittiaineistot: 8 TB/vuosi**

1 Gigabyte (1GB)
= 1000MB
CD album

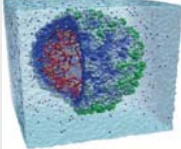
1 Terabyte (1TB)
= 1000GB
Word yearly
Book production

1 Petabyte (1PB)
= 1000TB
One LHC-
experiment
yearly data
production

1 Exabyte (1EB)
= 1000 PB
World yearly
information
production



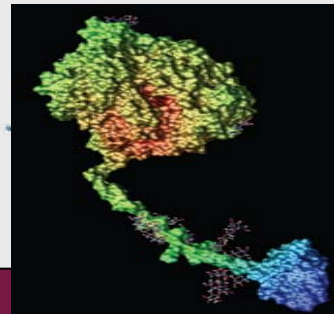
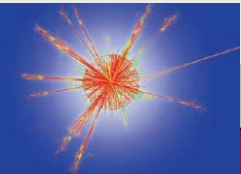
Haave paremmasta



VISIO: Tutkimuksen sähköiset tietoineistot ovat helposti ja avoimesti käytettävissä ja analysoitavissa, hyvin dokumentoituja, sopivasti suojattuja ja pysyvästi tallennettuja

Tämän aikaansaamiseksi tarvitaan kansallista tietopolitiikkaa, ohjausta ja linjauksia kysymyksiin:

- Omistajuus, IPR
- Tietosuojakysymykset
- Johdonmukainen tutkimusinfrastruktuurien rakentaminen
- Rahoitusratkaisut ja työnjako
- Ohjeet aineistojen jakamiseen, laatutyöhön, hallintaan ja käyttöön
- Hyvien käytäntöjen edistäminen
- Joustava kansainvälinen tekninen ratkaisu (=e-infrastruktuuri)
 - standardit, avoimet protokollat ja rajapinnat käytössä
 - käyttäjälle helppo pääsy, työkalut analysointiin ja visualisointiin
 - käyttäjien tunnistus, auktorisointi ja turvallisuus
 - joustavat ja toimivat työvuot



Tietopolitiikka säätelee

- **Huonosti hoidettuna se estää**

- aikaa, voimia ja luovuutta tuhraantuu neuvotteluihin
- aineistot ovat alikäytettyjä ja niitä unohtuu nurkkiin
- käytetään toissijaisia ja ilmaisaineistoja parhaiden asemasta

- **Hyvin toteutettuna se mahdollistaa**

- kaikki luovuus sisällölliseen työhön
- kotimaiset huippuaineistot arvonsa mukaiseen käyttöön
- lisäarvona korkeatasoisia T&K –tuloksia kansantaloudellisine vaikutuksineen
- hyväksi tunnettua seuraa tarjolla: USA, Kanada, ...

Tietopolitiikan kysymyksiä: Käyttöoikeuskysymykset ja avoin yhteiskäyttö

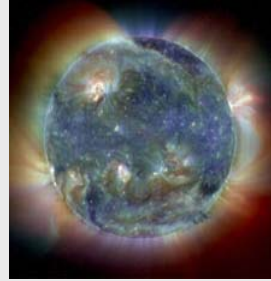
Suomessa ei ole yhtenäistä käytäntöä, mitä tulee tutkimuksen tietoaineistojen julkisuuteen, aineistojen yhteiskäyttöön ja avoimeen saatavuuteen. Monet kansainväliset tutkimussopimukset sisältävät myös tällaisen pyrkimyksen.

Julkisuuslainsäädännön, tilastolain ja muiden säädösten vaikutus aineistojen julkisuuteen muodostaa monimutkaisen vyyhdin, mikä on ollut esteenä aineistojen avoimelle yhteiskäytölle.

Tietokantojen ja muiden aineistojen julkistamista netissä on edistettävä, jotta niiden ympärille saadaan laajempaa tutkimusyhteistyötä ja viranomaiskäyttäjät saavat tarvitsemansa tiedot joustavasti. Tarvitaan aineistojen julkistamisen periaatteita koskeva ohjeistus, joka antaa säädöksistä selkeän tulkinnan ja sovittaa erilaiset tarpeet yhteen.



Mikä rajoittaa aineistojen tehokasta käyttöä?



- Vuorovaikutteisuuden puute
- Läpinäkyvyyden puute
- Yhteismitattomuus
- Aikaulottuvuuden heikko huomioiminen
- Korkea käyttökynnys

→ VAATIMUKSIA TOIMIVALLE RATKAISULLE:

- Käyttäjien ei tarvitse ymmärtää yksityiskohtaisesti tiedostomuotoja tai metatietokäytäntöjä
- Käyttäjien ei tarvitse ladata ja kerätä itselleen suuria aineistomääriä
- Käyttäjien ei tarvitse asentaa uusia ohjelmistoja
- Ratkaisun tulisi, aina kun mahdollista, noudattaa avoimia standardeja, protokollia ja rajapintoja yhteentoimivuuden kasvattamiseksi

Ratkaisun haasteet

- **Datamäärä**
- **Johdonmukaisten käytänteiden ja standardien puute**
- **Rahoittajien erilaiset vaatimukset datan hallinnasta ja jakelusta**
- **Ratkaisuja tuotetaan usein projektitasolla => ei kestävä ratkaisu**
- **Kansallista ratkaisua tarvitaan aineistojen elinkaaren hallintaan:**
 - tuottaminen, valinta, kokonaisuuksien koostaminen, pysyvä tunnus
 - metatiedot, versiot ja editiot; haku, säilytys
 - pääsy, tunnistus, turvallisuus
 - analysointi, synteesi, uudelleenkäyttö ja toistettavuus
 - maan rajat ylittävät työvuot

Arto Mustajoki: ”Tutkimusinfrastruktuurien tiekartta: harppaus kohti kansallista infrastruktuuripolitiikkaa”

”Yhteistyön tiivistäminen tutkimusinfrastruktuurien käytössä nousi paneelissa esiin. Yhteistyön vähäisyys ei sinällään ole millään tavalla vain suomalainen ongelma, vaan se näkyy myös kansainvälisesti. Paneeli halusi nostaa esiin erityisesti sellaiset hankkeet, joissa luodaan yhteistyölinkkejä. Toisin sanoen sellaisiin palveluihin, joihin kerätään aineistoja ja annetaan niitä laajasti käyttöön. Tällainen hanke on esimerkiksi Rekisteritutkimuksen tukikeskus ReTki”, Mustajoki sanoo.

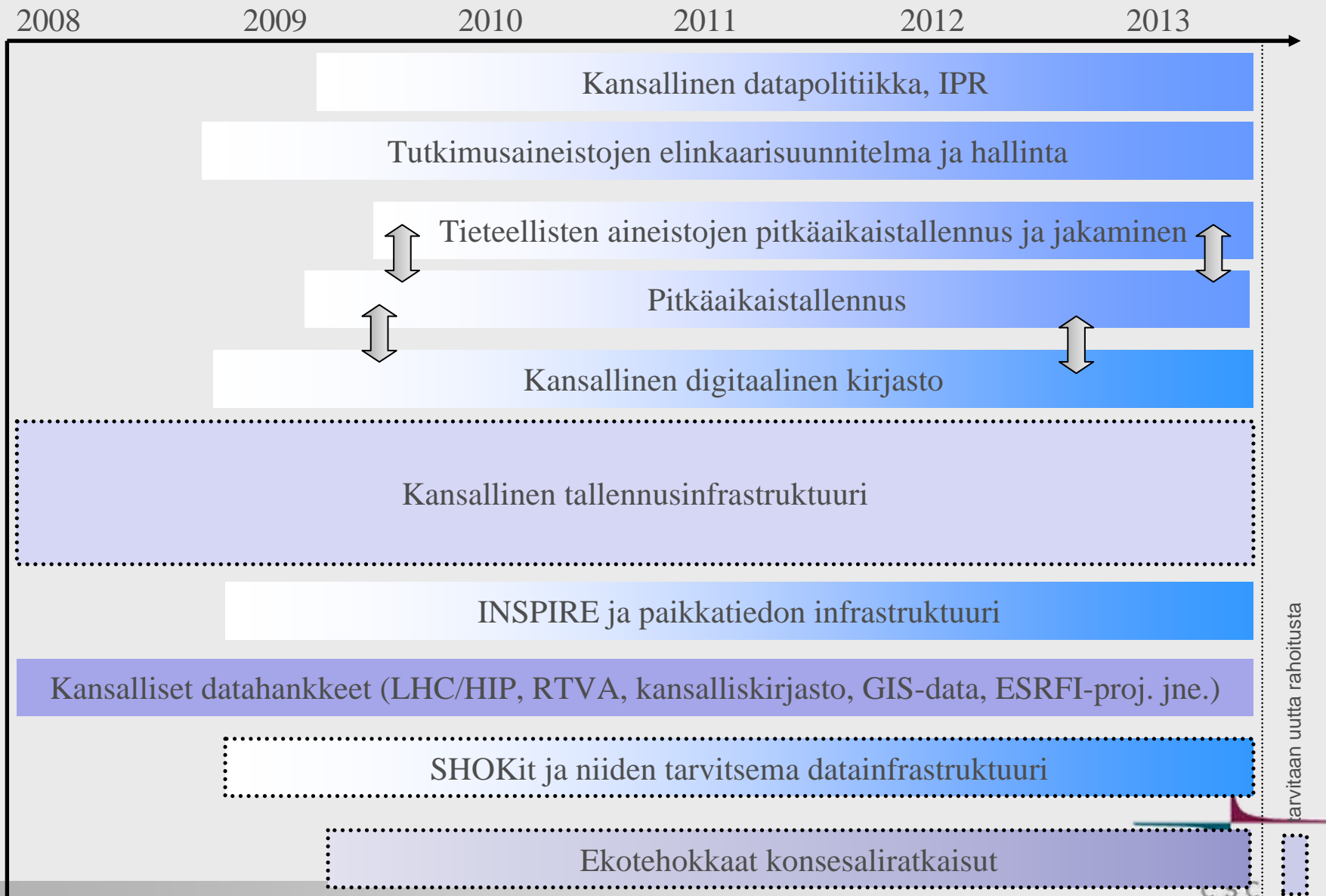
Tutkimuksen kannalta tärkeät aineistot pitäisi kiireellisesti saattaa digitaaliseen muotoon ja turvata alkuperäisaineistojen säilyminen tuleville sukupolville.

Myös tietovarantojen saatavuuteen, käyttäjäystävällisyyteen ja yhteiskäyttöisyyteen pitäisi kiinnittää huomiota. Käytännössä tämä tarkoittaa aineistopolitiikan kehittämistä nykyistä avoimempaan suuntaan, kansallisten aineistojen käytön maksullisuuden minimointia ja tutkijoiden liikkuvuuden kasvua.

(A propos 12.2.2009)



Kansallinen datan e-infrastruktuuri



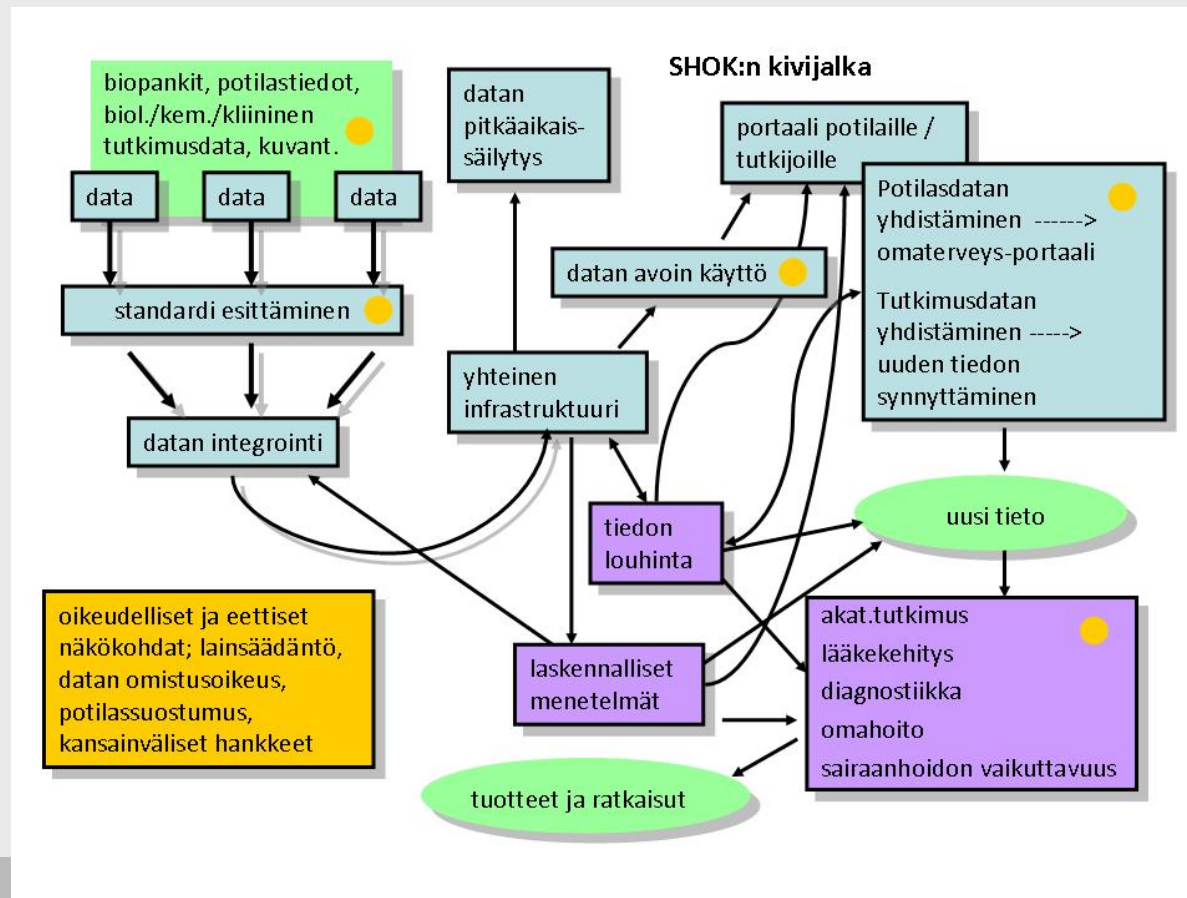
Esimerkkinä rekisteritutkimus:

Nykykäytäntöjen ongelmia

- **Aineiston käyttöön saaminen vie aikaa**
 - harvinaista ei ole 2 vuoden odotusaika lupaprosessin alusta tietojen saantiin; lisäaikaa kuluu korjauskierroksiin
- **Tietosuoja vaarantuu eri vaiheissa (henkilötiedot tunnistusteellisina)**
 - kirjepostin kuljetuksen aikana
 - tietojen käytön aikana: henkilötietojen käyttö jää tutkijan vastuulle
- **Aineiston anonymisointi (Tilastolaki)**
 - saatavat tiedot epätarkempia kuin tutkimuskysymyksiin vastaaminen edellyttäisi
 - otos ei (aina) riitä järkevään analyysiin
 - anonymisointi kallista ja virhealtista
- **Aineistojen hinnoittelukäytännöt vaihtelevia, suuri hintahaarukka**

SHOKien tutkimusaineistotarpeet

SHOK:n tutkimus- ja tuotekehitysprojekteissa käytettävä aineisto on alunperin hajallaan useissa lähteissä. Jos tieto kerätään yhteen, kuvaillaan standardien mukaisesti, integroidaan ja luodaan yhteinen infrastruktuuri, jonka avulla aineistoa voidaan analysoida helposti, syntyy uutta arvokasta tietämystä, ja aineisto on tehokkaasti käytettävissä tuotteiden ja ratkaisujen kehittämisessä.



Johtopäätös: tiedettä tukisi, jos

- **Julkisin varoin tuotettuja aineistoja olisi tarjolla avoimesti ja ajantasaisina**
 - Metadata, katselu, lataus yms.
 - Kaikille, ei vain tutkimus ja koulutustarkoituksiin (tukee taloutta, säästytään työläiltä rajanvedoilta)
 - Säädeltäköön sitä mikä ei voi olla julkista eikä päinvastoin
- **Lisäksi tutkijoille tulisi olla**
 - Muita aineisto-, ohjelmisto-, laskenta- ja rajapintapalveluita keskitetyllä ylläpidolla
 - Julkisen rahoituksen ehdoissa määritetty sitoutuminen avoimeen tietopolitiikkaan

OECD

- **OECD Principles and Guidelines for Access to Research Data from Public Funding (2007)**
- research data: numero-, teksti-, kuva- ja äänitallenteet

e-IRG
e-Infrastructure
Reflection Group



e-IRG
e-Infrastructure
Reflection Group

<http://www.e-irg.eu>

e-Infrastructures Roadmap 2007

Published by e-IRG



OPEN ARCHIVES

European Strategy Forum
on Research Infrastructures

ESFRI

EUROPEAN ROADMAP
FOR RESEARCH
INFRASTRUCTURES

Report 2006

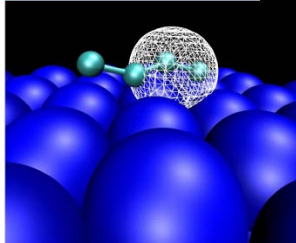
digitalpreservationEurope



EU-hankkeita

2007

European Computational
Science: the Lincei Initiative
"from computer to scientific
excellence"



European Science Foundation
19/02/2007



Towards a European e-Infrastructure
for e-Science Digital Repositories

planets



CASPAR

Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval



**COUNCIL OF
THE EUROPEAN UNION**



Council Conclusions on scientific information in the digital age: access, dissemination and preservation

*2832nd COMPETITIVENESS (Internal market, Industry and Research) Council meeting
Brussels, 22 and 23 November 2007*

- reinforce national strategies and structures for access to and preservation and dissemination of scientific information, tackling organisational, legal, technical and financial issues;
- enhance the co-ordination between Member States, large research institutions and funding bodies on access, preservation and dissemination policies and practices;
- maximise access for researchers and students to scientific publications, in particular by improving public procurement practices in relation to scientific information; this could include exchanging information on these practices and increasing the transparency of the contractual terms of "big deals", and exploring the possibilities for funding bodies, research institutions and scientific publishers from different Member States to work together in order to achieve economies of scale and efficient use of public funds by demand aggregation;
- ensure the long term preservation of scientific information - including publications and data - and pay due attention to scientific information in national information preservation strategies;



Bryssel 5.3.2009 KOM(2009) 108 lopullinen:

KOMISSION TIEDONANTO EUROOPAN PARLAMENTILLE, NEUVOSTOLLE, EUROOPAN TALOUS- JA SOSIAALIKOMITEALLE SEKÄ ALUEIDEN KOMITEALLE

TIETO- JA VIESTINTÄTEKNIikka TIETEEN PALVELUKSESSA

4.3. Tieteellisen tiedon saatavuuden parantaminen

Eurooppalaisissa ja kansallisissa e-infrastruktuureissa on otettava huomioon datakeskeisen tieteen myötä esiin nousevat uudet haasteet. Tätä varten Euroopan on luotava johdonmukainen ja hallittu tieteellisen tiedon varastojen ekosysteemi. Euroopassa on laadittava yhdenmukaiset periaatteet tieteellisen tiedon saatavuuden parantamiseksi (käyttämällä pohjana esimerkiksi ESFRI:n tieteellisestä datasta laatiman lausunnon linjauksia, komission tiedonantoa *Tieteellinen tieto digitaaliaikana: tiedon saatavuus, levittäminen ja säilyttäminen*³¹ sekä vuonna 2008 käynnistettyä seitsemännen puiteohjelman *Open Access*-pilottihanketta³²).

Jäsenvaltioita ja tiedeyhteisöjä pyydetään lisäämään investointeja tieteellisen datan infrastruktuureihin ja edistämään tiedonvaihtoa parhaista toimintatavoista.

Komissio vahvistaa seitsemännessä puiteohjelmassa katalyyttisiä investointeja tieteellisen datan infrastruktuuriin tukeakseen saatavuuteen ja säilyttämiseen liittyvien periaatteiden kehittämistä.



Suosittelut toimenpiteet Euroopassa:

- **Luotava johdonmukainen ja hallittu tieteellisen tiedon varastojen ekosysteemi.**
- **Laadittava yhdenmukaiset periaatteet tieteellisen tiedon saatavuuden parantamiseksi**

Perustana on käytettävä tutkimusalan keskeisten toimijoiden (esim. EMBL, ESA, ECMWF, CERN) sekä korkeakoululaitoksen ja kirjastojen jo kartoittamia ratkaisuja; ESFRIn suosituksia ja esimerkiksi INSPIRE-direktiivin tuella luotavaa paikkatietoinfrastruktuuria.

Tietoaineistoasian edistäminen

- **Tutkimuksen tietoaineistoja on Suomessa käsitelty vuosien mittaan eri osa-alueisiin (esim. tutkimusdata, tietokannat, tilastot) keskittyneissä työryhmäselvityksissä, mutta konkreettiset toimenpiteet ovat jääneet vähiin**
- **Yhteiskuntatieteellisen tietoarkiston perustaminen 1999.**
- **Kansallisarkiston aloite 2008**
- **Helsingin yliopiston työryhmä tutkimusaineistojen saatavuudesta 2008**
- **CSC:n selvityshanke 2009**

SUOSITUKSIA SUOMESSA

- **Suomen kansallisten tutkimusinfrastruktuurien kartoitus vuonna 2008: suositus kansallisten rekisterien käytettävyyden ja aineistojen saatavuuden parantamiseksi ja käyttäjän kustannusten alentamiseksi tarvittaessa ao. lainsäädäntöä muuttamalla.**
- **Tutkimus- ja innovaationeuvosto: on tärkeää edistää kansallisten rekistereiden ja erilaisten aineistojen käytettävyyttä ja saatavuutta sekä vaikuttaa käyttökustannuksien kohtuullistamiseen. tehostaa kansallisten tietovarantojen käyttöä.**
- **Suomen hallituksen politiikkariihen kannanotto 24.2.2009: useita toimia osaamisen ja tiedon luonnin ja hyödyntämisen parantamiseksi.**
- **Opetusministeriö kokoaa kansallisen poikkihallinnollisen selvityshankkeen kartoittamaan ja koordinoimaan julkisin varoin luotujen sähköisten tietoaineistojen ja tietovarantojen hyödyntämisen tehostamista.**

Kansallinen poikkihallinnollinen selvityshanke kartoittamaan ja koordinoimaan julkisin varoin luotujen sähköisten tietoaineistojen ja tietovarantojen hyödyntämisen tehostamista

Hankkeen ohjausryhmän tehtävät:

- 1) ohjata kansallista tutkimuksen tietoaineistoja koskevaa selvitystyötä ja muodostaa kokonaiskäsitys Suomen tilanteesta ja verrata sitä kansainvälisiin suosituksiin ja muiden maiden järjestelyihin
- 2) tehdä opetusministeriölle ja muille alan toimijoille ehdotuksia ja suosituksia myös jo ohjausryhmän toimiaikana tarpeellisista toimenpiteistä tutkimustietojen hyödynnettävyyden parantamiseksi
- 3) laatia kansallinen suunnitelma (tiekartta) tutkimustietojen saatavuuden ja säilytyksen kehittämiseksi huomioiden e-infrastruktuuriratkaisut sekä laatia ehdotus kansallisesta vastuunjaosta, yhteistyöstä ja koordinaatiosta eri toimijoiden kesken.

Ohjausryhmän määräaika on 31.12.2010.

Ohjausryhmän kokoonpano

Puheenjohtaja: kansleri Ilkka Niiniluoto, Helsingin yliopisto

Jäsenet:

- neuvotteleva virkamies Ralf Ekebom, sosiaali- ja terveysministeriö
- johtaja Pirjo-Leena Forsström, CSC - Tieteen tietotekniikan keskus Oy
- IT-johtaja Antti Holmroos, valtiovarainministeriö
- ylijohdaja Jarmo Hyrkkö, Tilastokeskus
- erikoissuunnittelija Väinö Malin, ympäristöministeriö
- ylijohdaja, tutkimus Riitta Mustonen, Suomen Akatemia
- pääjohtaja Jussi Nuorteva, Kansallisarkisto
- teollisuusneuvos Paula Nybergh, työ- ja elinkeinoministeriö
- tutkimusjohtaja Mikko Peltonen, maa- ja metsätalousministeriö
- ylijohdaja Pekka Plathan, liikenne- ja viestintäministeriö
- dosentti Marjut Salokannel, Helsingin yliopisto
- neuvotteleva virkamies Kirsti Vallinheimo, valtiovarainministeriö
- johtaja Leena Vestala, opetusministeriö
- alivaltiosihteeri Vesa Vihriälä, valtioneuvoston kanslia

Ohjausryhmän pysyviksi asiantuntijoiksi nimetään pääsuunnittelija Tuomas Parkkari tutkimus- ja innovaationeuvostosta ja erikoistutkija Markku Suvanen opetusministeriöstä.

Ohjausryhmän sihteerinä toimivat Satu Tissari ja Minna Laine CSC:stä.

Ohjausryhmän työtä tukee CSC:n selvityshanke ja sen kokoama asiantuntijaryhmä puheenjohtajanaan Pirjo-Leena Forsström. Lisäksi ohjausryhmä voi kuulla muita tahoja ja muita ulkopuolisia asiantuntijoita.

Selvitystyön eteneminen:

- **Ohjausryhmän ensimmäinen kokous** **kesäkuu 2009**
 - asiantuntijoiden valinta
 - tarkempi aikataulutus
- **Kansallinen kysely** **kesä 2009**
 - tilanne- ja tarvekartoitus
- **Asiantuntijoiden kuuleminen** **alkaen elokuu 2009**
 - asia/tieteenalakohtaiset paneelit
- **Kansainvälisen tilanteen selvitys** **syksy 2009**