Tutkimusdatan saatavuus ja aineistojen yhdistäminen yhteiskuntatieteissä

Sami Borg

STKS / Seminaari Avoin data -levällään? Tiedon louhinta, avoin data ja niihin liittyvät oikeudelliset kysymykset

Helsinki 22.1.2014

YHTEISKUNTATIETEELLINEN TIETOARKISTO
FINLANDS SAMHÄLLSVETENSKAPLIGA DATAARKIV
FINNISH SOCIAL SCIENCE DATA ARCHIVE

Esityksen rakenne

- I Datan saatavuus ja avaaminen yhteiskuntatieteissä
- II Datan yhdistämisestä yhteiskuntatieteissä
- Skenaarioita ja kysymyksiä

Datan saatavuus ja avaaminen yhteiskuntatieteissä: yleistä I

- **Aineistojen ja keruumahdollisuuksien moninaisuus**
- **Tutkimusta varten koottu data, muu julkisin varoin kertyvä data, muu data**
- **Datan keruuta, käyttöä ja säilyttämistä koskee joukko lakeja, jotka ohjeistavat (ja rajoittavat) myös tietojen yhdistelyä (Suomessa Henkilötietolaki, Tekijänoikeuslaki, Arkistolaki, Tilastolaki, Laki lääketieteellisestä tutkimuksesta jne.)**
- **Yhteiskuntatieteissä tutkimusta varten kootut aineistot kerätään usein itse, tai korkeintaan pienissä ryhmissä**

Datan saatavuus ja avaaminen yhteiskuntatieteissä: yleistä II

- Julkisin varoin tuotetaan runsaasti dataa, jotka sisältävät usein tietoja viranomaisrekistereistä
- Avattu viranomaisdata ei ole henkilötason ns. mikrodataa, vaan aluetasoista tai muuta kuin henkilödataa
- Internet on nopeasti laajeneva, valtava aineistolähde erityisesti KY-alan tutkijoille ja alan aineistojen muille käyttäjille
- Yhteiskuntatieteissä data koskee yleensä tavalla tai toisella ihmisiä, joilla on pääsääntöisesti oikeus määrätä itseään koskevien tietojen keruusta ja käytöstä
- Yhteiskuntatieteellinen tutkimus edellyttää hyvin usein henkilötason aineistoja riittävän luotettavien johtopäätösten tekemiseksi tieteellisesti hyväksyttäviä selityksiä varten

Infrastruktuureja ja avaamistoimia

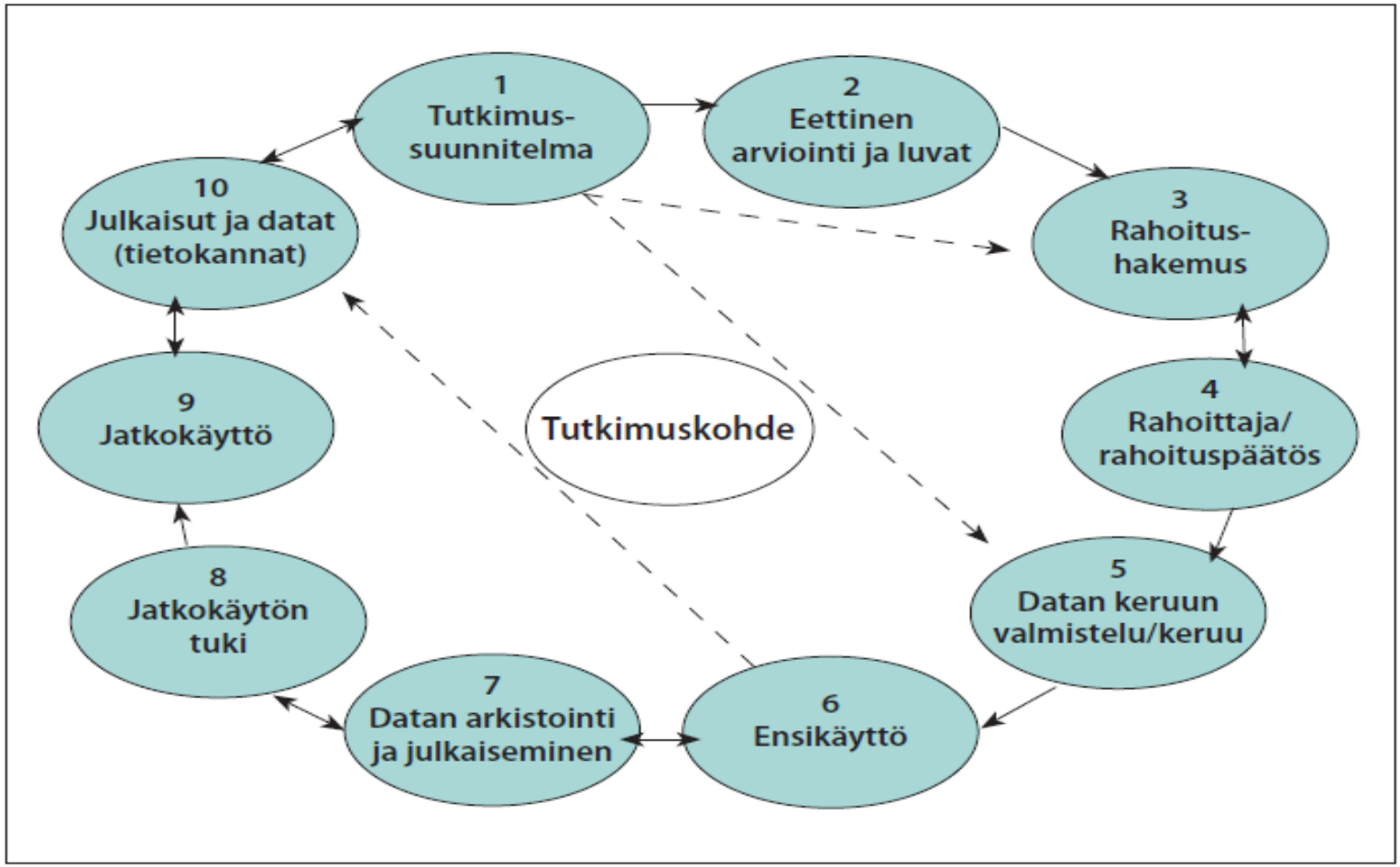
- Internetin, tutkimusinfrastruktuurien ja datapolitiikkojen kehittyminen on edistänyt tutkimusdatan avointa saatavuutta
- Suomeen Yhteiskuntatieteellinen tietoarkisto 1999-
- 2000-luvun alussa Suomen Akatemialta (SA) suositus tallentaa SA:n rahoittamissa hankkeissa kertyvä yhteiskuntatieteellinen aineisto tietoarkistoon
- OECD:n 2006 datasuositus julkisrahoitteisen tutkimusdatan avaamiseksi + lukuisat muut avaamissuositukset (myös EU)
- CSC:n johtamat hankkeet (Tutkimuksen tietoaineistot ja TTA) ovat edistäneet yleisesti datan avaamiseen liittyviä toimia; TTA tarjoaa uusia välineitä datan avaamiseen. TTA jatka vuodesta 2014 alkaen akronyymin ATT alla (Avoimen tieteen tiekartta)
- 2010-luvun alussa SA otti käyttöön pakollisen aineistonhallintasuunnitelman liitettäväksi kaikkien tieteenalojen rahoitushakemuksiin
- SA:lla tarkentuneet suositukset tutkimusdatan avaamiselle KY-aloilla (myös CLARIN) + suositus avoimesta julkaisemisesta

Tutkimusdatan avoin saatavuus: käyttäjänäkökulma

- **Tietosisällön laatu → käytön yleinen mielekkyys, soveltuvuus uusiin käyttötarkoituksiin**
- **Yleinen tekninen käytettävyys / käytön tehokkuus: koneluettavuus, yhteensopivuus**
- **Sisällön käytettävyys: dokumentointi, tarkistaminen, muokkaus, korjaaminen → virheettömyys, eheys, tarkkuus, läpinäkyvyys**
- **Saatavuuden helppous: helppo saavutettavuus (esim. www), maksuttomuus**
- **Käytön esteettömyys: käyttörajoituksettomuus tai korkeintaan vähäisiä käyttörajoituksia; kuitenkin käyttölisenssit ja käyttömahdollisuus lähdeviittausvelvoittein**
- **Yhdistettävyys ja linkitettävyys (datan harmonisointi, avainmuuttujat jne.)**
- **Open data: innovaatiot ja kaupallinen hyödynnettävyys**

Yhteiskuntatieteissä tutkimusdatan avointa saatavuutta ja käyttömahdollisuuksia rajoittavat/mahdollistavat useat seikat, joita voidaan tarkastella tutkimusaineiston elinkaaren perspektiivistä

Yksi tutkimusdatan elinkaarimalli / yhteiskuntatieteet



Lähde: Sami Borg & Arja Kuula (2007). Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari. Valmisteluraportti OECD:n datasuosituksen toimeenpanomahdollisuuksista Suomessa. Tampere: FSD.

Tutkimussuunnitelma ja datan avaaminen

- **Tutkimussuunnitelman ja tutkimusaineiston suunnittelun synkronointi: joskus aineiston täsmällinen suunnittelu vasta tutkimussuunnitelman laatimisen jälkeen**
- **Yhdistämisen kannalta olennaisia: tutkimusongelmat sekä tutkimusmenetelmät ja –aineisto → mittaustaso ja muuttujien operationalisointi**
- **Aineiston luokittelu-, harmonisointi- ja yhdistämis-tarpeet ensikäytön aikana (linkkimuuttujat)**
- **Aineiston yhdistämistarpeet myöhemmin (linkkimuuttujat)**

Eettinen arviointi ja luvat: ennakkoarviointi I

- **Eettinen arviointi koskee ihmistieteissä kaikkia tutkimussuunnitelmia mutta aiheuttaa laajempia toimenpiteitä vain tietyissä tapauksissa**

- **”Eettisellä ennakkoarvioinnilla tarkoitetaan tutkimussuunnitelman arviointia tieteenalakohtaisten eettisten käytänteiden mukaisesti painottuen tutkimuksesta tai sen tuloksista tutkittavalle mahdollisesti koituvan haitan ennakointiin.”**

- **”Ihmistieteellisessä tutkimuksessa eettiset kysymykset painottuvat tutkijan ja tutkittavan kohtaamiseen, johon voi sisältyä ennakoimattomia tekijöitä. Tutkija vastaa aina itse tutkimuksensa eettisistä ja moraalisista ratkaisuista”**

Eettinen arviointi ja luvat: ennakkoarviointi II

- **TENK 2009: Humanistisen, yhteiskuntatieteellisen ja käyttäytymistieteellisen tutkimuksen eettiset periaatteet ja ehdotus eettisen ennakkoarvioinnin järjestämiseksi: → ei yleistä etukäteisarviointia Suomeen**
- **Ihmistieteisiin luettavaa tutkimusta koskevat eettiset periaatteet jaetaan kolmeen osa-alueeseen:**

1. **Tutkittavan itsemääräämisoikeuden kunnioittaminen**
2. **Vahingoittamisen välttäminen**
3. **Yksityisyys ja tietosuoja**

TENK: Yksityisyys ja tietosuoja

- **”Yksityisyyden suojaa koskevat tutkimuseettiset periaatteet jaetaan kolmeen osaan: 1)tutkimusaineiston suojaaminen ja luottamuksellisuus, 2) tutkimusaineiston säilyttäminen tai hävittäminen ja 3) tutkimusjulkaisut. Periaatteiden lähtökohtana on pyrkimys sovittaa yhteen luottamuksellisuuden ja tieteen avoimuuden periaate. ”**
- **”Yksityisyyden suojaa koskevia periaatteita ei sovelleta yleisesti saatavilla oleviin julkisiin aineistoihin ja julkistettuihin tietoihin, jotka voivat koskea yksittäisiä henkilöitä ja heidän toimiaan politiikan, elinkeinoelämän, viranomaistoiminnan ja kulttuurin parissa. Yksityisyyttä koskevia ohjeita ja tietosuojaperiaatteita on kuitenkin noudatettava oikeudenistuntoja ja tuomioistuinten päätöksiä koskevien asiakirjojen osalta. ”**
- **”Tunnisteellisten aineistojen käsittelystä säädetään henkilötietolaissa (523/1999). Lain 3 § :n mukaan henkilötiedolla tarkoitetaan "kaikenlaisia luonnollista henkilöä taikka hänen ominaisuuksiaan tai elinolosuhteitaan kuvaavia merkintöjä, jotka voidaan tunnistaa häntä tai hänen perhettään tai hänen kanssaan yhteisessä taloudessa eläviä koskeviksi.”**
- **Lähde: TENK http://www.tenk.fi/sites/tenk.fi/files/eettisetperiaatteet.pdf**

TENK: Yksityisyys ja tietosuoja (jatkuu…)

- **Tunnisteellisuudesta puhuttaessa merkittävintä on se, voiko yksittäistä henkilöä tunnistaa tiedoista helposti ja kohtuuttomitta kustannuksitta. Tutkimusaineiston tunnisteet on perinteisesti jaettu käsitteellisesti suoriin eli yksilöiviin ja epäsuoriin tunnistetietoihin.**

- **Suoria tunnistetietoja ovat nimi, osoite, henkilötunnus, syntymäaika sekä ihmisen ääni ja kuva.**

- **Epäsuoria tunnisteita ovat esimerkiksi kotipaikkakunta ja asuinalue, koulutus, työpaikka ja perheen koostumus.**

- **Lähde: TENK**
  http://www.tenk.fi/sites/tenk.fi/files/eettisetperiaatteet.pdf

Tutkimuksen eettinen ennakkoarviointi

- **”Eettisessä ennakkoarvioinnissa tarkastellaan aineistonkeruun suunnitelmaa, tutkimuksen toteutustapaa, tutkittavien informointia sekä aineiston käsittelyn ja säilyttämisen suunnitelmaa riskien ja vahingon välttämisen näkökulmasta. Arvioinnissa punnitaan tutkittaville tutkimukseen osallistumisesta mahdollisesti koituvia haittoja ja vahinkoja suhteessa tutkimuksella tavoiteltavaan tietoarvoon. Arvioinnin ohjeellisena lähtökohtana ovat ihmistieteiden eettiset periaatteet (tutkittavien itsemääräämisoikeus, vahingoittamisen välttäminen, yksityisyys ja tietosuoja)”**

Lähde: http://www.tenk.fi/sites/tenk.fi/files/eettisetperiaatteet.pdf

- **Eettistä ennakkoarviointia harjoittavat kuusi alueellista eettistä toimikuntaa sairaanhoitopiireissä sekä yliopistojen ja muiden tutkimusorganisaatioiden eettiset toimikunnat → lausunnot tutkimussuunnitelmista**

- **Suuri osa ennakkoarvioinnin piiriin kuuluvasta tutkimusdatasta hävitetään ensikäyttövaiheen jälkeen → ei avointa dataa**

Rahoitushakemus ja -päätös

- Tutkimussuunnitelma
- Mahdollisesti lausunto eettiseltä toimikunnalta
- Aineistonhallintasuunnitelma pakollinen Suomen Akatemialle esitettävissä hankkeissa
- Tutkimusrahoittajan datapolitiikka aineiston avaamisessa vaikuttaa saatavuuteen

Aineiston keruu

- Kerättävän aineiston perussisältö lyödään lukkoon
- Keruu- ja tallennusvaihe määrittää yhdistelymahdollisuudet
- Tutkittavien informointi ja tutkimussuostumus vaikuttavat keskeisesti aineiston avaamiseen ja jatkokäyttämahdollisuuksiin
- → avaamista ja yhdistettävyyttä tukevat suostumukset

Aineiston arkistointi ja avaaminen

- **”Akatemia suosittelee myös, että sen rahoittamat hankkeet luovuttavat kokoamansa yhteiskuntatieteellisen aineiston Yhteiskuntatieteellisen tietoarkiston www.fsd.uta.fi käyttöön. Samoin suositellaan, että hankkeissa luodut kieliaineistot saatetaan muiden tutkijoiden käyttöön FIN-CLARIN-järjestelmän kautta.”**
- **”Akatemia kehottaa julkaisemaan tutkimustulokset avoimissa tiedejulkaisuissa silloin, kun alalla on valittavissa perinteisiin lehtiin verrattuna vähintään samantasoisia sähköisessä muodossa olevia tieteellisiä julkaisuja tai tallentamaan artikkelit avoimesti saatavilla olevaan sähköiseen arkistoon.”**

Lähde: http://www.aka.fi/fi/A/Tutkijalle/Hakeminen/Hakuohjeet_/Yleiset-hakuohjeet/

- **CSC:n palvelut: IDA, KATA, REMS, AVAA Ks. http://www.tdata.fi/**
- **Tutkijoiden omat väylät datan avaamiselle**

Tutkimusaineiston tekijänoikeudet

- Tekijänoikeuden kohteena voi olla kirjallinen tai taiteellinen teos.
- Olennaista tekijänoikeudellisen suojan saamiseksi on se, että tuote tai tuotos ylittää teoskynnyksen.
- Teossuojan saamiseksi teoksen on oltava tekijänsä luovan työn omaperäinen tulos (olisiko kukaan muu samaan työhön ryhtyessään voinut päätyä samanlaiseen lopputulokseen?)
- Yhteiskuntatieteelliseen tietoarkistoon tallennetut aineistot ovat kysely- ja haastatteluaineistoja. Suuri osa empiirisistä tutkimusaineistoista ei ylitä teoskynnystä, mutta tutkimusaineistoihin voi sisältyä myös tekijänoikeudellisesti suojattuja osia.
- Jo aineistonkeruuta suunniteltaessa on tärkeää sopia aineiston käyttöoikeuksista koskien sekä alkuperäistä tutkimusta että sen jälkeistä aikaa.
- Lähde: Tietoarkiston ylläpitämä Tutkimusaineistojen tiedonhallinnan käsikirja http://www.fsd.uta.fi/tiedonhallinta/osa2.html#keruun_suunnittelu

Tutkimusdatan arkistointi / omistus,- hallinta- ja tekijänoikeudet

- Julkisrahoitteisen tutkimusdatan omistaa yleensä tutkimus- tai tiedontuottajaorganisaatio
- Tekijänoikeudet ovat tekijöillä; tutkimusdata kannattaa julkaista, jolloin datalle on annettava viitetiedot, jotka ilmaisevat myös tekijyyden
- Hallintaoikeus voidaan ja se tulisi ymmärtää laajasti. Liian usein tutkimusdata jää yksinomaan sen keränneen tutkimusryhmän käyttöön. Kerääjien tulisi avata datan saattamalla se avoimeen jatkokäyttöön tarkoituksenmukaista reittiä pitkin.
- Jatkokäyttöön avaaminen ei tarkoita tekijyyden ja hallintaoikeuden menettämistä. Jatkokäyttäjillä on viittausvelvollisuus ja avoin käyttömahdollisuus tuo mukanaan lisää viittauksia tutkimukseen ja sen dataan.
- Rekisteritutkimuksen tietosuojaopas tutkijoille ja tietopyyntöjä käsitteleville rekisterinpitäjille

http://www.tietosuoja.fi/uploads/mst1e.pdf

- Ks. myös Salokannel, Marjut (2013): Tekijänoikeus ja tutkimuksen raaka-aineet. Uuden teknologian haasteet tekijänoikeudellisesti suojatun materiaalin tutkimuskäytölle. Koneen säätiö, 2013.

Datan arkistointi ja luovutus jatkokäyttöön: esimerkkinä tietoarkisto

- **Arkistointisopimus**
  http://www.fsd.uta.fi/fi/lomakkeet/Arkistointisopimus.pdf
- **Käyttölupahakemus**
  http://www.fsd.uta.fi/fi/lomakkeet/Kayttolupahakemus.pdf
- **Käyttöehtositoumus**
  http://www.fsd.uta.fi/fi/lomakkeet/

Jatkokäyttö + jatkokäytön julkaisut

- Datan rikastaminen
- Uudet julkaisut tutkimusdatasta
- Datan ja julkaisujen linkittäminen (pysyvät tunnisteet tutkimusdatalle ja tutkimusjulkaisuille)

Datan avaaminen ja yhdistämismahdollisuudet: avoin data ja toimijaroolit

Taulukko 2.1. Esimerkkejä toimijoiden rooleista suhteessa dataan.

Tallentaja	raakadatan kerääminen ja tallentaminen
Jalostaja	raakadatan käsittely ja muokkaaminen
Aggregaattori	datan yhdistely ja koostaminen eri lähteistä
Harmonisoija	eri lähteistä tulevien tietojen yhdenmukaistaminen ja yhteismitallistaminen (samalta näyttävä asia myös tarkoittaa samaa)
Päivittäjä	tietojen päivittäminen
Julkaisija	datan julkaiseminen
Rekisterinpitäjä	datavarannon ylläpitovastuu
Sovelluskehittäjä datan loppukäyttäjänä	datan hyödyntäminen osana palvelua
Tulkitsija datan loppu-käyttäjänä	datan tulkitseminen, esim. tutkija, yritys tai demokra-tiaaktivisti
Dataan pohjautuvien palveluiden (ks. luku 2) käyttäjät	ihmiset yritykset ja organisaatiot, jotka käyttävät avoimen datan päälle tehtyjä sovelluksia ja tulkintoja

Poikola, Antti, Kola, Petri & Hintikka, Kari A.: *Julkinen data: johdatus tietovarantojen avaamiseen*. Helsinki: Liikenne- ja viestintäministeriö, 2010

Data yhdistäminen: esimerkkinä surveydata

- Miten ja miksi dataa yhdistetään?
- Tutkittavien ilmiöalueiden liittäminen toisiinsa (selittävät ja selitettävät tekijät)
- Havaintoyksikköjen yhdistäminen aineistokokonaisuudeksi: tyypillisesti tilastollinen käyttötarkoitus
- Linkki- tai avainmuuttujat mahdollistavat tietojen yhdistämisen havaintoyksikkökohtaisesti (havaintoyksikön tunnistemuuttuja, henkilötunnus, kunta jne.)
- Aineistoa voidaan täydentää lisäämällä muuttujia ja/tai havaintoyksikköjä
- Yksilötason aineistoon voidaan lisätä yksilötason muuttujia tai aggregoituja kontekstuaalitietoja esim. kuntamuuttujan avulla
- Periaatteessa esim. aluetason aineistoon voidaan lisätä yksilötason aineistoista aggregoituja lisätietoja

Analyysitasot ja tietojen yhdistäminen

- **Henkilöaineistoissa havaintoyksikkönä on yksittäinen henkilö. Tietoja kerätään ja yhdistetään aineistosta toiseen pääsääntöisesti tutkittavan henkilön omalla suostumuksella (pl. Viranomaisvaltuudet ja -rekisterit).**
- **Vaikka henkilötason aineistot olisi anomymisoitu, niihin voi silti yhdistää muita kuin pelkästään henkilöä itseään koskevia tietoja (esim. kunta, asuinalue jne.).**
- **Henkilö voi itse avata itseään koskevia tietoja, joskus myös ehkä harkitsemattomasti (internet).**

Avoimen, yhdistettävissä olevan datan käyttö: skenaarioita

- Uudet, innovatiiviset tutkimuskysymykset ja niiden ratkaiseminen
- Tiedon visualisointi yleistyy, ehkä hyvässä ja pahassa
- Tiedonkeruu halpenee ja uudet aineistonkeruutavat korvaavat vähitellen vanhoja, jo vaikeuksissa olevia tiedonkeruutapoja
- Villi länsi? Yleistyvätkö akateemiset karjapaimenet? (suuret datamassat ja vauhtisokeus saattavat hämärtää tutkimuseettisten näkökohtien huomioon ottamista)
- Internetin kaupalliset toimijat pyrkivät rahastamaan uusiin kysymyksiin ja lähestymistapoihin soveltuvilla tiedoilla
- Käytetään dataa jota saadaan eikä dataa, jollainen sopisi parhaiten tutkimusongelmiin

(Esim. ristipainehypoteesi: ”Politiikkaan osallistutaan aktiivisemmin alueilla, jotka ovat poliittisesti yhdenmukaisia kuin poliittisesti erimielisillä, jakautuneilla alueilla” → suuri ekologisen virhepäätelmän riski)

Linked data (”yhdistettävissä oleva data”)

- **”Linked data is an approach to publication of data on the web. It is a set of best practices to enable systems to use the Web to connect related data that wasn't previously linked, or lower the barriers to linking data currently linked using other methods. “**
- **“The approach [@@ref] recommends use of HTTP URIs to name the entities and concepts so that consumers of the data can lookup those URIs to get more information, including links to other related URIs. “**
- **“RDF [@@ref] provides a standard for the representation of the information that describes those entities and concepts, and is returned by dereferencing the URIs”**
- **Lähde: http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html**

Linked Open Data

★	**Available on the web (whatever format) but with an open licence, to be Open Data**
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context