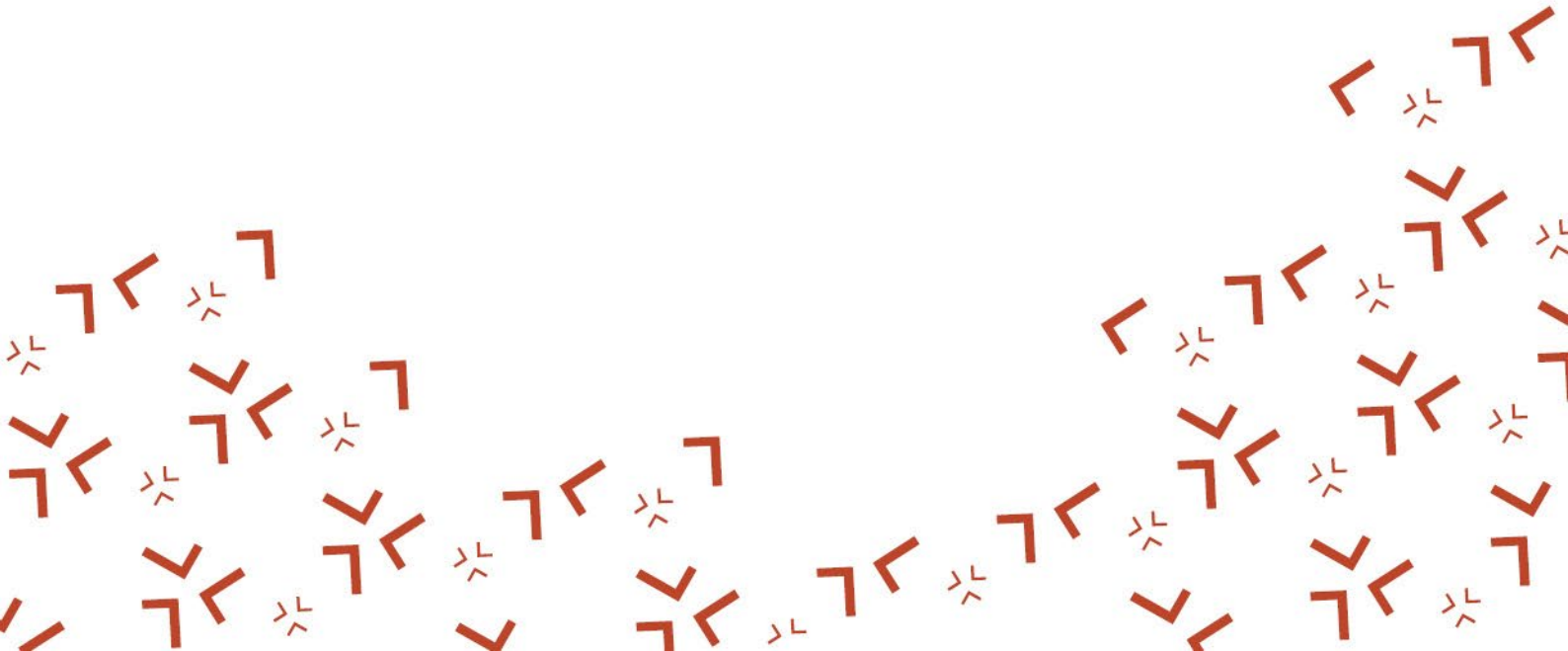

Heli Väätäjä

Avoimen datan merkitys ja mahdollisuudet

TAUSTAA

Avoimesta datasta ja sen merkittävyydestä puhutaan paljon, mutta mitä sillä tarkoitetaan, mistä sitä saa, kuka sitä tuottaa, miten sitä voi käyttää sekä miksi se on merkityksellistä myös ammattikorkeakouluissa?

Tässä artikkelissa kuvataan yhteenvetoa Tiedolla johtamisen asiantuntija YAMK -koulutuksen Data-analytiikka ja avoin data kehittämisen työkaluna -opintojaksolla keväällä 2021 opittua avoimesta datasta ja sen hyödyntämisestä. Myös henkilökunnan oli mahdollista osallistua avoimen datan asiantuntijan, Mika Honkasen, luentoihin. Opintojakson oppimistehtävässä osa opiskelijoista etsi ja analysoi ja muutama myös yhdisti avointa dataa oman organisaationsa dataan.

AVOIN DATA JA SEN MERKITYS

Avoimen datan määritelmä

Avoin data määritellään seuraavasti (Avoimen datan opas):

> *”Avoin data on digitaalisessa muodossa olevaa informaatiota, joka on kaikkien vapaasti käytettävissä mihin tahansa käyttöön, kunhan sen alkuperäinen lähde mainitaan”.*

Tarkempi määritelmä asettaa avoimelle datalle seuraavia ehtoja:

1. Saatavuus ja saavutettavuus – Datan tulee olla saatavilla kokonaan ja kohtuullisilla luovutuskustannuksilla. Datan tulisi mielellään olla ladattavissa internetistä. Datan tulee olla myös käyttökelpoisessa ja muokattavassa muodossa.
2. Uudelleenjakelu ja uudelleenkäyttö – Data tulee tarjota sellaisin ehdoin, että se mahdollistaa uudelleen käytön ja uudelleen jakelun, mukaan lukien yhdistämisen muihin datasetteihin.

3. Maailmanlaajuinen osallistuminen – Kaikilla tulee olla mahdollisuus käyttää, uudelleen käyttää sekä uudelleen jakaa dataa. Käytössä (lisensoinnilla) ei saa rajoittaa käyttökohteita, eikä henkilöitä tai ryhmiä asettaa eriarvoiseen asemaan. Esimerkiksi rajoittaminen ei-kaupalliseen käyttöön (estäen kaupallisen käytön), tai rajoitukset vain tiettyyn käyttöön, kuten opetukseen, eivät ole sallittuja.

Honkanen määritteli avointa dataa luennollaan seuraavasti: Avoin tarkoittaa avoimen datan tapauksessa, että tiedossa on avoin käyttölupa eli lisenssi (CC BY 4.0 tai CC0 1.0) eli avoin saatavuus ja vapaa käyttö on sallittu. Data puolestaan tarkoittaa digitaalisesti tallennettua, koneluettavaa informaatiota, joka koostuu merkeistä ja symboleista (esim. taulukoita, tekstiä, kuvia, karttoja, videoita, äänitiedostoja tmv.).

Avoimen datan lisenssit eli käyttöluvat

Avoimen datan käyttölupia esitellään tässä kaksi maailmanlaajuisesti käytettävää lisenssiä.
CC0 1.0 Yleismaailmallinen (CC0 1.0) -lisenssi ei rajoita hyödyntämistä mitenkään, mutta lähteen maininta lisää uskottavuutta ja on hyvän tavan mukaista. Lisenssin antaja luopuu lainsäädännön rajoissa kaikista tekijänoikeuksista työhönsä. Suositeltu käyttötarkoitus lisenssille on kuvailutiedot (metatiedot) ja osa tietoaineistoista. Tätä lisenssiä voi käyttää esim. tutkimusdatan julkaisemiseen. Lisenssi tarjoaa hyvin vapaat käyttöoikeudet: 1. jakaa ja kopioida, 2. käyttää kaupallisesti, 3. julkaista muokattuna, 4. voit jakaa eri lisenssillä, 5. tekijän nimen voi jättää mainitsematta. (Creative Commons Suomi)

Creative Commons Nimeä 4.0 Kansainvälinen julkinen -lisenssin eli CC BY 4.0 (Creative Commons Attribution 4.0 International Public License) -lisenssin yhteydessä on oltava teksti ”Tämä teos on lisensoitu Creative Commons Nimeä 4.0 Kansainvälinen -lisenssillä”, jossa on oltava linkki kyseiselle selitesivulle. Lisenssiä käytettäessä lähde on mainittava, tarjottava linkki lisenssiin ja mainittava, jos on tehty muutoksia. Suositeltu käyttötarkoitus on erilaiset tietoaineistot, teksti, kuvat ja muu media. Lisenssi tarjoaa muutoin samat mahdollisuudet kuin CC0 1.0 lisenssi, mutta tekijän nimi pitää mainita. Lisenssiä suositellaan muun muassa tutkimuksen julkaisuun (Open Access) sekä avoimen oppimateriaalin julkaisuun (OER). Lisäksi lisenssiin voidaan lisätä vaatimus jakamisesta Share Alike eli SA merkinnällä (CC BY-SA 4.0). (Creative Commons Suomi)

Datan ja avoimen datan hyödyntämisen tilanne

Honkanen kuvasi luennollaan, miten maailman arvokkaimmat yritykset toimivat datalla – sen keräämisellä ja hyödyntämisellä. Esimerkkinä Honkanen käytti Netflixiä, jossa jatkuvasti kerätään tietoa käyttäjien käyttäytymisestä palvelussa. Tämä tukee asiakkaiden tarpeiden parempaa tunnistamista. Suomessa esimerkiksi Koneen liikevaihdosta tulee jo 50% tulee palveluista, joissa hyödynnetään kerättyä dataa.

Julkinen sektori tuottaa Honkasen mukaan koko ajan valtavan määrän dataa. Dataa kerätään paljon, mutta sitä analysoivia työntekijöitä on vähän. Tämän vuoksi dataa hyödynnetään vähän eli alikäytetään. Julkisuuslain mukaan kaikki julkisen hallinnon keräämä ja luoma tieto on julkista, jos sen salaamiseen ei ole laissa määriteltyä perustetta. Myös ammattikorkeakoulut tuottavat monenlaista tietoa ja dataa, joka on julkista, avointa ja jota voidaan datasta riippuen anonymisoitunakin jakaa.

Avoimen datan tarjonta on toistaiseksi painottunut julkiseen sektoriin – valtionhallintoon ja suurimpiin kuntiin. Yksityinen sektori on sitä vastoin jakanut melko vähän dataa avoimesti. Data halutaan pitää yksityisenä, kun se liittyy omiin tuotteisiin, työntekijöihin ja asiakastietoon eikä alustatalouden mahdollisuuksia osata vielä hyödyntää. Dataa kuitenkin käytetään yrityksissä tuotekehitystyössä ja asiakaskokemuksen kehittämisessä. Usein avointa dataa yhdistetään yrityksen omiin aineistoihin, asiakkaan tietoihin tai muihin aineistoihin. Dataa ja datayhdistelmiä voidaan hyödyntää innovatiivisesti uusien sovelluksien ja palveluiden kehittämisessä.

Suomessa avoimen datan kehittäminen kansallisesti alkoi vuonna 2011. Suomen hallitus julisti tuolloin periaatepäätöksen julkishallinnon digitaalisten aineistojen saatavuudesta: ”tietoaineistojen tulee olla avoimesti saatavilla ja uudelleenkäytettävissä yhtenäisin, selkein ja kaikille tasapuolisin ehdoin, pääsääntöisesti maksutta.” Maanmittauslaitos oli ensimmäisten joukossa avaamassa aineistoja, avaten maastotietoja vuonna 2012.

”Datan arvo kasvaa, mitä enemmän sitä käytetään”, totesi Honkanen luennollaan. Julkisella sektorilla avoin data lisää demokratiaa ja hallinnon läpinäkyvyyttä, vähentäen väärinkäytöksiä ja korruptiota ja toisaalta paljastaa mitä parannettavaa toiminnassa on. Avoin data myös mahdollistaa organisaation toiminnan ja tehostumisen läpinäkyvyyden kautta ja samalla uusien ideoiden syntyminen mahdollistuu. Kansalaisten aktiivisuus ja osallistuminen mahdollistuu datan avoimuuden myötä ja muuttuu passiivisesta tiedon vastaanottajasta aktiiviseksi tiedon käyttäjäksi ja hyödyntäjäksi. Myös uusien palveluiden kehittäminen mahdollistuu. Honkasen mukaan tästä voivat hyötyä erityisesti pienet ja keskisuuret yritykset.

AVOIMEN DATAN JAKAMINEN JA HYÖDYNTÄMINEN

Avoimen datan jakaminen

Honkanen esitteli kolme avoimen datan jakelutapaa:

- tiedostona
- latauspalvelun kautta
- rajapinnan (API) avulla.

Tiedosto on yksinkertaisin, staattisin (korkeintaan päivittäin päivittyvä) ja helpoin käsitellä datan hyödyntäjän näkökulmasta. Tiedostoja voi yleensä käsitellä suoraan toimisto-ohjelmilla. Tyypillisiä esimerkkejä ovat postinumerot, asukasmäärät, ja erilaiset tilastot. Dataa hyödynnetään laskennassa ja visualisoinnissa. Esimerkiksi avoindata.fi -palvelu toimii parhaiten tiedostojen jakeluna.

Latauspalvelun kautta tarjottavien tiedostojen tiedot voivat päivittyä usein. Datan käsittely on vaikeampaa kuin tiedostojen käyttö. Tiedostojen lataaminen on usein helppoa, ja niiden käsittely saattaa onnistua valmisohjelmistoilla. Esimerkkeinä Honkanen mainitsi Suomen kartan päälle tehdyt tietokerrokset, kuten kierrätyspisteet, bussipysäkit, luonnonsuojelualueet. Tietoa hyödynnetään visualisoinneissa, laskennassa, ja tietojen yhdistämisessä.

Trendinä on siirtyä tiedostoista rajapintojen hyödyntämiseen. Rajapinnan kautta tarjottava tieto päivittyy tyypillisesti useita kertoja päivässä. Tietoa hakee ihmisen sijaan toinen ohjelma ja rajapinta toimii kyselyvastaus -ketjuna. Rajapinnan avulla saadun datan käsittely on vaikein kolmesta jakelutavasta ja dataa voidaan joutua suodattamaan. Usein rajapinnan käyttöön tarvitaan ohjelmointitaitoa, mutta käyttö voi onnistua myös ilman ohjelmointia, jos rajapinta on suunniteltu helppokäyttöiseksi. Esimerkkejä avatusta datasta ovat julkisen liikenteen kulkuneuvojen reaaliaikaiset sijainnit, sääennusteet ja yritystiedot. Rajapinnan kautta jaettavaa avointa dataa hyödynnetään esimerkiksi erilaisissa sovelluksissa, visualisoinneissa, analytiikassa, ja koneoppimisessa. Monet älypuhelimen sovellukset hyödyntävä avointa dataa rajapinnan kautta.

Avoimeen dataan liittyvä lainsäädäntö

Keskeinen avoimeen dataan ja tietoon liittyvä lainsäädäntö liittyy Suomen perustuslakiin (731/1999), julkisuuslakiin, EU:n tietosuoja-asetukseen, sekä avoimen datan direktiiviin (PSI, 2019/1024/EU) sekä tekijän- ja teollisuusoikeuksiin. Perustuslaki on kaiken lainsäädännön ja julkisen vallan käytön perusta. Julkisuusperiaatteen mukaisesti viranomaisten asiakirjat ovat julkisia, jollei Laissa viranomaisten toiminnan julkisuudesta (607/2016) tai muussa laissa erikseen toisin säädetä. Esimerkiksi valtionhallinnosta noin 1% on salaista tietoa, muu avointa.

Eniten tällä hetkellä keskustelussa on sekä EU:ssa että Suomessa henkilötietojen suoja suhteessa avoimuuteen eli avoimuuden perusperiaate ja vastapainona henkilötietojen suoja. Henkilötietoja ovat kaikki tiedot, jotka liittyvät tunnistettuun tai tunnistettavissa olevaan henkilöön. Henkilötietoja ovat siis tiedot, joiden perusteella henkilö voidaan tunnistaa suoraan tai välillisesti. Välillisesti tunnistaminen voi tapahtua yhdistämällä yksittäinen tieto johonkin toiseen tietoon, joka mahdollistaa tunnistamisen. Henkilötietoja on käsiteltävä lainmukaisesti, asianmukaisesti, läpinäkyvästi, luottamuksellisesti, turvallisesti, vain tiettyä, nimenomaista ja laillista tarkoitusta varten sekä kerättävä vain tarpeellinen määrä henkilötietojen käsittelyn tarkoitukseen nähden. Tietosuoja suojelee henkilötietoja. Honkasen mukaan henkilötietojen määrittely onkin asia, jota mietitään eniten ennen data avaamisprosessia ja joiden määrittely on usein vaikeaa.

Avoimen datan laatu

Avoimen datan hyödynnettävyyteen, jatkokäyttöön sekä automaattiseen hyödyntämiseen vaikuttaa datan julkaiseminen avoimessa ja yleisesti tunnetussa tiedostomuodossa koneluettavassa muodossa. Avoin ja rakenteellinen muoto ja datan sisällön

ymmärrettävyys helpottavat sen käyttöä. Honkanen esitteli avoimen datan viiden tähden mallin (Tim Berners-Lee, 2009) yleisimpänä datan laatua kuvaavana mallina.

Viiden tähden mallin alimmalla, yhden tähden tasolla, data on internetissä saatavilla missä tahansa tiedostomuodossa ja käyttölisenssi on avoin. Tyypillinen esimerkki on PDF tiedosto aineistosta. Toisella, kahden tähden tasolla tietoaineisto on saatavilla avoimessa tiedostomuodossa, esim Microsoft Excel, Power Point tai Word tiedostona. Microsoft Excel onkin edelleen suosittu tapa käsitellä dataa. Kolmen tähden tasolla tietoaineisto on saatavilla avoimessa tiedostomuodossa, esim. CSV -tiedostona. Neljännellä tasolla datassa on yksilöllinen ja elinikäinen tunniste, jonka avulla datan sisälle pystyy viittaamaan suoraan eri kohtiin. Esim. SPARQL on W3C-standardoitu kyselykieli RDF tietokantaan. Ylimmällä eli viiden tähden tasolla tietoaineistossa on linkkejä sen ulkopuolisiin aineistoihin. Tämä mahdollistaa verkostomaisen avoimen datan selailun. Aineistot muodostavat kokonaisuuden ja tietoaineistojen välisen liikkumisen.

Avoimen datan lähteistä

Tilanteesta ja datan tyypistä riippuen dataa voi löytyä eri lähteistä. Honkasen mukaan tyypillisiä lähteitä ovat

1. datakatalogit, jotka keräävät tietyn alueet tai toimialan tietoaineistojen kuvailutietoja
2. yksittäisen julkaisija oma datakatalogi
3. yksittäiset raportit ja aineistot
4. verkkosivuilla olevat materiaalit, joita voidaan kerätä automaattisestikoneluettavaa muotoon sopivalla skreippaajalla (tiedot haravoivalla skriptillä)
5. joukkoistamalla eli kysymällä/pyytämällä suurilta ihmisjoukoilta, kun valmista dataa ei ole olemassa.

Honkanen ohjeisti aloittamaan datan etsimisen suurista datakatalogeista, joihin on kerätty tuhansia tietolähteitä ja julkaisijoita. Usein datakatalogeissa on vain datan kuvailutiedot eli metadata. Sopivan aineiston löytyessä linkin avulla pääsee joko datatiedostoon tai kyselyrajapintaan.

Suomessa erityisesti suuret kunnat ovat olleet aktiivisimpia datan avaajia. 6Aika -hankkeen puitteissa Helsinki, Espoo, Vantaa, Tampere, Turku ja Oulu ovat avanneet aineistojaan omissa portaaleissaan. Esimerkiksi pääkaupunkiseudun portaalissa HRI:ssä (Helsinki Region Infoshare) oli lokakuussa 2021 638 data-aineistoa, 284 sovellusta ja 168 rajapintaa. Samaan aikaan Oulun kaupungin portaalissa on 82 data-aineistoa, 11 sovellusta ja 27 rajapintaa (Oulun kaupungin dataportaali). Aineistot kattavat seuraavia alueita: asuminen, hallinto ja päätöksenteko, talous ja verotus, kartat, terveys- ja sosiaalipalvelut, kulttuuri ja vapaa-aika, väestö, liikenne ja matkailu, ympäristö ja luonto, rakennettu ympäristö sekä opetus ja koulutus.

Avoindata.fi on portaali, joka kokoaa yhteen Suomessa julkaistuja tietoaineistoja. Suomen ympäristökeskus SYKE on avannut portaaliin poronhoitoalueiden laidunluokituksen, jotta tietoa voisi hyödyntää porolaidunten käytön suunnittelussa (Suomen ympäristökeskus, 2021). SYKE on avannut myös esimerkiksi virtavesien lohikalakantojen esiintyvyysaineiston, jota voidaan hyödyntää suunnittelun tukena lohikalojen elinmahdollisuuksien parantamiseksi ja kantojen vahvistamiseksi (Suomen ympäristökeskus, 2020). Portaalin kautta saatavilla olevia aineistoja voivat hyödyntää yhtä lailla kansalaiset kuin yritykset ja julkishallinnon eri toimijat.

Vaikka yritysten avaamia aineistoja on vielä vähän, Fingrid on ensimmäisenä dataa avannut sähköverkkoyhtiö EU:ssa. Suomessa tutkimusaineistoja julkaistaan tutkimus-, opetus- ja opiskelukäyttöön mm. Tietoarkistossa (Tietoarkisto). Tilastokeskus tarjoaa myös monipuolisia avoimia tilastoaineistoja hyödynnettäväksi päätöksentekoon ja tutkimukseen.

Etsin -palvelun avulla voi etsiä tutkimusaineistoja ja niiden metatietoja Fairdata -palveluista (Etsin). Tutkimusaineistojen kuvailu- eli metatiedot ovat julkisia, mutta aineiston omistaja voi päättää aineiston käyttöoikeudesta. Tutkimusaineistoja on myös monissa kansainvälisissä portaaleissa, kuten CERNin ylläpitämässä monialaisessa ZENODO palvelussa (ZENODO). Tutkimusaineistojen osalta noudatetaan FAIR-periaatteita, joihin myös Suomen opetus- ja kulttuuriministeriö on sitoutunut. FAIR-periaatteiden mukaan datan on oltava löydettävissä (Findable), saavutettavissa (Accessible), yhteentoimivaa (Interoperable) ja uudelleenkäytettävissä (Re-usable) (FAIRDATA).

Avoimia data-aineistoja on siis monipuolisesti saatavilla hyödynnettäväksi monenlaiseen tarkoitukseen opetuksesta uusien TKI-projektien ja liiketoimintamallien ideointiin, toteutukseen ja kehittämiseen.

AVOIMEN DATAN MAHDOLLISUUKSIA AMMATTIKORKEAKOULUISSA

Avoin data tarjoaa monia mahdollisuuksia ammattikorkeakoululle TKI-toiminnassa, opetuksessa ja muussa kehittämistoiminnassa. Erilaisiin avoimiin tietoaineistoihin ja lähteisiin tutustuminen avaa oivalluksille tilaa hyötykäytön mahdollisuuksista. Toisaalta ammattikorkeakoulukin voi avata dataa omasta toiminnastaan opetukseen, TKI-toimintaan sekä muuhun toimintaan liittyen. Oppilaitoksilla on erityisen tärkeä rooli avoimuuden periaatteiden levittämisessä yhteiskuntaan myös opiskelijoiden saavuttamien tietojen ja taitojen, toiminnan läpinäkyvyyden sekä TKI-toiminnan käytäntöjen ja tulosten jakamisen kautta. Avoimen datan käyttöä ja jakamista voidaan lisätä ja rohkaista myös ammattikorkeakouluissa erilaisten ketterien kokeilujen ja niistä saatujen kokemusten sekä tulosten jakamisen avulla.

LÄHTEET

Avoimen datan opas. Mitä on avoin data? Digi- ja väestötietovirasto. Viitattu 13.10.2021 https://www.avoindata.fi/fi/opas/mita-on-avoin-data

Berners-Lee, T. 2009. 5-star deployment model for open data. Viitattu 13.10.2021 https://5stardata.info/en/

Creative Commons Suomi. Tietoa lisensseistä. Viitattu 13.10.2021 https://creativecommons.fi/lisenssit/

Etsin. Fairdata.fi. Opetus- ja kulttuuriministeriö. Viitattu 13.10.2021 https://etsin.avointiede.fi/

FAIRDATA. FAIR-periaatteet. Opetus- ja kulttuuriministeriö. Viitattu 13.10.2021 https://www.fairdata.fi/tietoa-fairdatasta/fair-periaatteet/

Fingrid. Avoin data. Viitattu 13.10.2021 https://data.fingrid.fi/

Helsinki Region Infoshare. Viitattu 13.10.2021 https://hri.fi/fi/

6Aika. Viitattu 13.10.2021 https://6aika.fi/

Oulun kaupungin dataportaali. Viitattu 13.10.21 https://data.ouka.fi/fi/

Suomen ympäristökeskus. 2021. Poronhoitoalueiden laidunluokitus. Viitattu 13.10.2021 https://www.avoindata.fi/data/fi/dataset/poronhoitoalueiden-laidunluokitus

Suomen ympäristökeskus. 2020. Virtavesien lohikalakannat. Viitattu 13.10.2021 https://www.avoindata.fi/data/fi/dataset/virtavesien-lohikalakannat

Tilastokeskus. Avoin data. Viitattu 13.10.2021 https://www.stat.fi/org/avoindata/index.html

Tietoarkisto. Viitattu 13.10.2021 https://www.fsd.tuni.fi/fi/

ZENODO. Cern Data Centre. Viitattu 13.10.2021 https://www.zenodo.org/