Eikö olisi hienoa, että kaikki julkishallinnon hallussa oleva tieto, mikä on julkista, olisi myös helposti löydettävissä ja saatavissa yhdestä paikasta?

6. Avoimen datan infrastruktuuri

Tässä luvussa esittelemme idean datakatalogeista sekä datan avaamisessa avustavista poikkihallinnollisista toimijoista, joita kutsutaa tässä yhteydessä clearing house -organisaatioiksi.

Poikkihallinnolliset toimijat ovat datanavauksen konsultoijia, jotka auttavat yhdistämään osapuolten intressejä. Datakatalogi konkreettisena verkkosivustona on tällaisen poikkihallinnollisen tukiorganisaation sekä datan julkaisijoille että uudelleenkäyttäjille tarjoama palvelu. Poikkihallinnollinen toimija voi olla kansallisella tasolla oma tietohallintovirastonsa, kuten Iso-Britannian Office of Public Sector Information (OPSI), tai seudullisesti tai kaupungin sisällä toimiva hallintokuntien rajat ylittävä tukiorganisaatio.

Datakatalogi on strukturoitu metadatarekisteri, johon on yhdistetty metadataa useampien julkisten organisaatioiden hallussa olevista aineistoista. Datakatalogit voivat olla kansallisia (mm. suomi.fi/datakatalogi ja data.gov.uk), seudullisia (Washington D.C. tai mahdollinen Helsinki Region Infoshare), kaupunkien ylläpitämiä (San Francisco) tai yksityisten tahojen ylläpitämiä (Sunlight Foundation - National datacatalog). Ideaalitapauksessa datakatalogin ylläpitäjä voi tarjota tukea datan julkaisijoille esimerkiksi liittyen lisensointiin tai ohjelmallisesti käsiteltäviin rajapintoihin ja formaatteihin. Datakatalogia ylläpitävä taho voi myös toimia kontaktina datan hyödyntäjiin organisoimalla esimerkiksi uudelleenkäyttöön kannustavia kilpailuja tai keräämällä kokemuksia, palautetta ja datan avauspyyntöjä.

Case: Kevyesti alkuun – Suomen ensimmäinen datakatalogi

Suomi.fi:n toiminta vuonna 2009 järjestetyn Suomen ensimmäisen Apps for Democracy -kilpailun yhteydessä on esimerkki nopeasta reagoinnista ja toimintaan ryhtymisestä. Kilpailussa kehitettiin palveluita julkisten koneluettavien tietolähteiden varaan. Tarvittiin paikka, johon koota linkit aineistoihin. Suomi.fi:n toimitus otti tehtäväkseen ylläpitää listaa tietolähteistä aivan tavallisella verkkosivulla Laatua verkkoon -projektin sivuston yhteydessä. Toteutus ei ollut tietolähteiden esillepanon näyttävyyden tai löydettävyyden kannalta optimaalinen, mutta se saatiin aikaiseksi lyhyellä varoitusajalla.

Datakatalogin esiaste (http://suomi.fi/datakatalogi) vaikutti enemmän kuin Suomi.fi:n toimittajat ehkä arvaavatkaan. Kilpailu

ja katalogi osoittautuivat tärkeiksi välineiksi datan avaamisen konkretisoimisessa. Nopea, kevyt ja hyvin ajoitettu reagoiminen tuotti ehkä suuremman tuloksen, kuin laajempi mutta hitaampi toiminta.

Sivusto ei ollut pelkästään lista tietolähteistä, vaan se oli signaali siitä, että hallinto tukee kilpailua ja haluaa avata datavarantojaan. Suomi.fi:n toimitus toimi välittäjänä, joka puhui sekä muiden hallinnon organisaatioiden että kilpailun järjestäjien kanssa. Ilman hallinnon tukea kilpailu olisi saanut erilaisen vireen. Sivuston julkistaminen on hyvä esimerkki siitä, minkälaisia kerrannaisvaikutuksia pienellä teolla voi olla avoimen datan ekosysteemin kehittymisen kannalta Suomessa.

6.1 Datakatalogi – kaikki julkinen data yhdeltä luukulta

Eikö olisi hienoa, että kaikki julkishallinnon hallussa oleva tieto, mikä on julkista, olisi myös helposti löydettävissä ja saatavissa yhdestä paikasta?

Katalogien avulla datavarantojen olemassaolo tehdään tunnetuksi potentiaalisille datan hyödyntäjille. Tarve julkisten datavarantojen löydettävyyden parantamiseksi on huomattu myös EU:n PSI-direktiiviä (jakso 3.5.1) seuranneen työryhmän suosituksissa: *"Käytännöllisiä hankkeita datavarantojen rekistereiden ja muun PSI-infrastruktuurin luomiseksi tulisi tukea kansallisesti ja myös rajoja ylittäen Euroopan tasolla"*. Suositustekstin mukaan datakatalogit, jotka sisältävät rakenteellisessa muodossa metatietoa sekä julkaistuista että toistaiseksi vielä julkaisemattomista hallinnon datavarannoista, ovat välttämättömiä hallinnon tietovarantojen uudelleenkäytön lisäämisessä, ja ne tuottavat samalla merkittäviä hyötyjä dataa hallinnoiville organisaatioille itselleen.

Vuoden 2009 aikana tämä ajatus yhden luukun periaatteella toimivista julkisen datan katalogeista löi maailmalla läpi kenties Obaman hallinnon esimerkkiä seuraten. Yhdysvaltain hallinnon data.gov julkaistiin toukokuussa 2009 ja sen jälkeen julkaistiin maailmalla kymmeniä julkishallinnon datavarantoja listaavia ja jakelevia datakatalogeja valtioiden, kaupunkien ja kaupunkiseutujen tasolla (esim. San Francisco, Lontoo, New York). Monissa paikoissa, missä hallinto ei vielä ole datakatalogia tehnyt, ovat kansalaiset olleet aktiivisia, kuten Ruotsissa (opengov.se). Vuotta 2009 voidaan hyvällä syyllä kutsua kansainväliseksi julkishallinnon datavarantojen julkaisemisen vuodeksi.

Muutamaa poikkeusta lukuun ottamatta ennen vuotta 2009 julkaistut datakatalogit keskittyivät rajatummin esimerkiksi ainoastaan paikkatietoon, vaativat rekisteröitymistä tai olivat tarjolla rajatuille kohderyhmälle, kuten tutkijoille. Datakatalogien julkaisun lisäksi Yhdysvaltain ja Britannian hallitukset myös sitoutuivat strategisesti avoimen datan tuottamiseen (Digital Britain ja Open government directive).

Case: data.gov.uk

Toistaiseksi ehkä näyttävin hallinnon datakatalogi on 19.1.2010 julkisena beta-versiona avattu Ison-Britannian data.gov.uk. Tämä www:n kehittäjän Tim Berners–Leen johdolla kuudessa kuukaudessa koottu sivusto päihittää USA:n data.govin mennen tullen. Julkaisuhetkellä Britannian katalogissa on noin 3000 tietoaineistoa, mikä on kolme kertaa enemmän kuin data.gov:ssa.

Datan määrän ohella Iso-Britannian katalogin sisältöä voidaan pitää myös kiinnostavampana. USA:n data.gov:ia on sen julkaisusta asti kritisoitu siitä, että siellä on vain sellaisia aineistoja, jotka eivät vahingossakaan herätä poliittista keskustelua. Britannina katalogista sen sijaan löytyy mm. armeijassa tapahtuneiden kuolemantapausten tilastoja ja muuta läpinäkyvyyden mielessä kiinnostavaa aineistoa. Britannian katalogi on myös rakennettu tukemaan koneluettavuutta ja semanttisen webin ominaisuuksia. Suoraan katalogin etusivulta pääsee sivulle, jonka kautta voi selaimessa tehdä tietokantahakuja SPARQL nimisellä kyselykielellä. Myös hakutuloksia – esimerkiksi brittiläisiä kouluja – voi tutkia selaimessa ilman että omalle koneelle tarvitsee ladata mitään. Katalogissa käytetyn ratkaisun merkitys on, että siinä käytettyyn teknologiaan perehtyminen onnistuu erittäin nopeasti.

6.1.1 Datakatalogien tekninen yhteensopivuus

On luontevaa, että valtionhallinnolla on oma virallinen datakataloginsa, jonka rinnalla toimii paikallisia ja aihealuekohtaisia katalogeja. Julkisen datan tuottajina seudulliset ja kuntatason toimijat ovat erittäin merkittävässä asemassa, sillä suuri osa uudelleenkäytön kannalta kiinnostavasta datasta löytyy niiden organisaatioilta. Eri maissa datakatalogien yleistyminen on saanut julkisuutta yleensä korkean profiilin kansallisen datakatalogin myötä. Kehitys on kuitenkin usein käynnistynyt jo aiemmin paikallisen tason datakatalogeilla. Esimerkiksi Kanadassa on jo lukuisia eri kaupunkien omia katalogeja, mutta kansallinen sivusto on siellä vasta kehitteillä.

Virallisten tahojen ohella myös yksityiset tahot ovat perustaneet omia julkisten datavarantojen katalogeja, kuten Sunlight Foundation Yhdysvalloissa ja Ruotsin opengov.se. Näitä katalogeja on perustettu toisaalta sen takia, kun julkista katalogia ei ole ollut ja toisaalta sen takia, että ohjelmoijat ja kansalaiset saattavat vaihtaa ajatuksiaan vapautuneemmin omalla datayhteisösivulla, jonka ylläpito on viranomaisista riippumatonta.

Kansallinen taso ei missään tapauksessa ole julkishallinnon datavarantojen hyödyntämisessä korkein tarkastelutaso. Erityisesti EU:n tasolla on kiinnostusta jäsenmaiden väliseen yhteistyöhön ja sitä kautta monikansallisiin

julkishallinnon datakatalogeihin ja ajan myötä myös aineistojen harmonisoimiseen niiden käytettävyyden lisäämiseksi.

Iso-Britannian datakatalogi on toteutettu kokonaan avoimen lähdekoodin ratkaisuilla. Julkaisujärjestelmänä on Drupal ja aineistojen metadatan hallinta on toteutettu Open Knowledge Foundationin vuodesta 2006 lähtien kehittämällä CKAN-alustalla. Tähän mennessä kaupallisista järjestelmätoimittajista ainoastaan Microsoft on julkaissut julkishallinnon datakatalogeja varten oman tuotteensa *Open Government Data Initiative*, joka on otettu ensimmäisenä käyttöön Edmontonin kaupungissa.

Datakatalogien yhteensopivuus liittyy oikeastaan keskeisesti kysymykseen siitä, mitä metadataa julkishallinnon datalähteistä halutaan kerätä. Tähän ei ole vielä olemassa valmiita laajasti hyväksyttyjä standardeja. Niiden valmistumista ei myöskään kannata jäädä odottelemaan, sillä ne muotoutuvat olemassaolevien ja syntyvien katalogien ja käytäntöjen kautta. Tässä vaiheessa riittää, että huolehditaan Suomeen syntyvien katalogien keskinäisestä yhteensopivuudesta ja yhteentoimivuudesta merkittävimmän eurooppalaisen katalogin eli data.gov.uk:n kanssa.

6.2 Taustatyö datakatalogin takana

Päällepäin edellä kuvatut datakatalogit ovat parhaimmillaan yhden luukun periaatteella toimivia sivustoja, joiden kautta pääsee kiinni julkiseen tietoon. Pintaa syvemmältä datakatalogit edustavat ns. clearing house -ajattelua. Clearing house -organisaatio on datan käyttäjien ja tuottajien välissä toimiva taho, jonka tehtävänä on kerätä, säilyttää ja levittää informaatiota, metadataa ja dataa. Clearing house on finanssitoiminnasta yleistynyt termi välittäjäorganisaatiolle. Suomeksi se voisi olla vaikka selvitystoimisto, mutta käytämme tässä oppaassa englanninkielistä termiä.

Datan hyödyntäjille ja jatkojalostajille tällainen clearing house -organisaatio saattaisi näyttäytyä hyvin ylläpidettynä katalogina julkishallinnon tietovarantoihin. Yhdestä paikasta löytyisi kaikki informaatio, mikä lain mukaan on julkista. Datakatalogin taustalla toimiva clearing house -organisaatio selvittäisi työkseen lukuisten muiden organisaatioiden kanssa teknisiä, juridisia ja organisatorisia kiemuroita sekä harmonisoisi ja kuvailisi yhdenmukaisesti eri lähteistä tulevaa dataa. Se voisi myös hoitaa julkisen tietokannan ylläpidon ja kehittämisen, niin että lukuisat organisaatiot, jotka nykyisin ylläpitävät ns. operatiivisia tietokantoja omia tarkoituksiaan varten voisivat ulkoistaa tämän toiminnon clearing house -organisaatiolle ja keskittyä informaation jalostamiseen.

Clearing house olisi siis vastuussa siitä, että raakadata olisi koneluettavassa muodossa saatavissa kaikille sitä tarvitseville ja siitä kiinnostuneille niin hallinnon sisällä kuin ulkopuolellakin. Sillä ei olisi mitään velvoitteita jalostaa dataa eikä

tuottaa ns. perusnäkymiä, joissa data on esitetty luettavassa muodossa. Nykyisin perusnäkymän tekeminen on viranomaisten velvollisuus ja koneluettavien rajapintojen tekeminen jää usein lapsipuolen asemaan. Raakadatan tarjoaminen rajapintoineen voitaisiin ulkoistaa clearing house -organisaatiolle ja viraston tehtäväksi jäisi edelleen vain se, mikä sillä on nykyäänkin: peruspalveluiden tuottaminen ja informaation jalostaminen.

Toki informaation jalostuksen myötä syntyy sellaista uutta tietoa, jota ei voida enää kutsua raakadataksi, mutta joka voidaan haluttaessa edelleen tarjota rajapintojen kautta maksutta eteenpäin. Tällöinkin virasto voisi ulkoistaa tämän jalostetun informaation tarjoamisen clearing house'lle.

6.2.1 Julkisen datan avaamispalvelu

Vaikka julkinen data on periaatteessa saatavilla, saattaa sen käytännön hyödynnettävyydessä ilmetä ongelmia, kun dataa halutaan kopioida, yhdistää muihin datalähteisiin, jalostaa tai julkaista uudelleen. Haasteet saattavat liittyä esimerkiksi maksullisuuteen, käyttöehtoihin tai formaatteihin, jossa julkinen informaatio on tarjolla (2.2 avoimen datan mittareita).

Datan hyödyntäjien kannalta olisi luontevaa voida ilmoittaa käytännön ongelmista keskitetysti yhdelle taholle, jolla olisi valtuuksia ja keinoja pyrkiä yhdessä datan tarjoajan kanssa parantamaan aineiston käytettävyyttä. Iso-Britanniassa edellä mainittu OPSI tarjoaa julkisen datan avaamispalvelua (Public Sector Information Unlocking Service), jonne voi ilmoittaa mm. aineistoista, jotka puuttuvat data.gov.uk katalogista tai ovat muutoin hankalasti käytettyjä.

Julkisen datan avaamispalvelua voisi verrata esimerkiksi kuluttajavalistuslautakuntaan, mutta ainakin Britanniassa se on brändätty positiivisempaan sävyyn. Palveluun jätetyt ilmoitukset ovat kaikille avoimia ja muut käyttäjät voivat kommentoida ja äänestää ehdotuksia ja niitä ratkotaan tärkeysjärjestyksessä.

6.3 Suomalaisia hankkeita

Julkisen datan infrastruktuurin kehittäminen on Suomessa juuri nyt erittäin ajankohtaista. Syksyllä 2010 valmistuvan kansallisen tietoyhteiskuntastrategian valmistelussa yhtenä kuudesta teemasta on esitetty *"Tiedon roolia yhteiskunnan perusinfrastruktuurina on tuettava"* (Otakantaa.fi 2010). Työ- ja elinkeinoministeriön tuoreessa 16.2.2010 julkaistussa käyttäjälähtöisen innovaatiopolitiikan toimenpideohjelmassa (TEM 2010) esitetään toimenpiteeksi: *"Arvioidaan sidosryhmien kanssa tarve perustaa kansallinen julkisen sektorin tiedon hyödyntämistä koskeva yhteyspiste, jonka tehtävänä olisi mm. jakaa tietoa, tarjota opastusta sekä verkottaa toimijoita."*

Kuten luvun alussa kerroimme, vuonna 2009 Suomi.fi otti ansiokkaasti pallon haltuunsa julkishallinnon datakatalogin kehittämisessä. Kevättalvella 2010

suomi.fi kehittää seuraavaa versiota datakatalogista. Tavoitteena on rakentaa virallinen paikka, mihin hallinnon edustajat voisivat ilmoittaa omia datavarantojaan. Valtion hallinnoiman katalogin ohella kehitetään parhaillaan myös kirjastot.fi:n tuella yhteisöllistä pääasiassa ohjelmoijille ja palveluideoijille suunnattua opengov.fi -sivustoa. Siellä datan hyödyntäjät voivat keskustella ja jakaa epämääräistäkin tietoa tyyliin, *"Helsingilläkin on varmaan olemassa rekisteri kaupungin istutetuista puista"*. Tavoitteena on kevyt sivusto, joka toimii yhteen muiden virallisempien katalogien kanssa. Pääkaupunkiseudun kuntien kesken on käynnistymässä seututiedon harmonisoimiseen ja maksuttomaan jakamiseen tähtäävä *Helsinki Region Infoshare* -hanke.

Vaikka yleiset julkisen datan katalogit ovat verrattain uusi ilmiö, on esimerkiksi paikkatiedoille ollut hakupalveluita käytössä jo useita vuosia. Maanmittauslaitos on parhaillaan ottamassa käyttöön uuden sukupolven paikkatietohakemisto.fi-palvelua, joka perustuu avoimen lähdekoodin GeoNetwork -ohjelmistoon. Paikkatietojen osalta INSPIRE-direktiivi on määritellyt paikkatietojen metadatan yhdenmukaisesti (GEMET-sanasto) ja kuvannut niitä varten kehitettävät kansalliset ja euroopanlaajuisen hakupalvelun. Direktiivin toimeenpanon yhteydessä on huomioitu organisatorisia ja hallinnollisia asioita, jotka eivät ole vain paikkatietosidonnaisia. Niitä voidaan soveltaa avoimen datan infrastruktuurin kehittämisessä.

Metadatan ja sanastojen määrittelyssä edellisessä luvussa esitetty Linked-data-lähestymistapa on eduksi. Julkaisemalla hallinnon datakatalogin RDF-muodossa kaikkea ei tarvitse määritellä etukäteen, vaan tarvittaessa uusia määrittelyitä voidaan ottaa käyttöön vaikka käsite kerrallaan. Aalto yliopiston semanttisen webin tutkimusryhmä kehittää parhaillaan semanttisen datakatalogin prototyyppiä. Myös käynnissä olevassa valtiollisen tason tietoarkkitehtuuriprojekstissa (VALTASA) noudatetaan semanttisen webin ajattelutapaa (VM 2009).