Verkossa toimii voimakas pienimmän yhteisen nimittäjän voima. Monikäyttöiset, yksinkertaiset ja helposti omaksuttavat ratkaisut, joilla voidaan kattaa laajoja sovellusalueita, saattavat levitä yllättävänkin laajaan käyttöön.

KUVA: diebmx / FLICR.com | CREATIVE COMMONS

5. Tekninen valmistautuminen

Tämä luku tarjoaa yleiskatsauksen avoimen datan tekniseen käsitteistöön. Näkemyksemme mukaan kaikkien datan avaamisen kanssa työskentelevien olisi hyvä ymmärtää avoimen datan teknologiset perusteet yleisellä tasolla. Toivomme teknologiaa selventävien määrittelyjen auttavan käytännön toiminnassa, kun sekä teknisen alan ammattilaiset että muiden alojen asiantuntijat ja päättäjät keskenään keskustelevat ja päättävät datan avaamisen politiikasta.

Aiemmissa luvuissa olemme opastaneet tekemään inventaarion organisaation tietovarannoista sekä tarkastelleet datan julkaisemista taloudellisesta ja juridisesta näkökulmasta. Nämä ovat toivon mukaan antaneet työkaluja päätöksentekoon. Seuraava vaihe on määritellä datan julkaisemisen tekniset puitteet, joilla on suuri merkitys avoimen datan hyödyntämiselle ja jatkoprosessoinnille. Jos datan hyödyntäminen on liian vaikeaa, jäävät avoimen datan mahdollisuudet toteutumatta.

Kuten juridiikkaluvussa listasimme tärkeimmät avoimeen dataan liittyvät lait ja säännökset, niin myös tähän alkuun on koottu ja kuvailtu kaikki tässä luvussa esiintyvät teknologiat ja standardit (Taulukko 5.1). Standardien, yhteyskäytäntöjen eli protokollien ja formaattien (datan esitysmuotojen) määrää ei kannata säikähtää. Niiden muistaminen tai syvällinen sisäistäminen ei ole tarpeellista, sillä listoihin voi myöhemminkin palata. Tässä yhteydessä niitä käytetään pikemminkin apuneuvoina, joiden kautta voidaan viitata datan julkaisuun liittyviin teknisiin ratkaisuvaihtoehtoihin. Lista ei myöskään pyri olemaan kattava, mutta se sisältää näkemyksemme mukaan kuitenkin tärkeimmät teknologiat.

Taulukko 5.1 Avoimen datan julkaisuun liittyviä teknologioita

Standardeja datan esittämisen ohjelmallisesti käsiteltävässä muodossa

- **XML:** Extensible Markup Language (XML) on yleiskäyttöinen merkintäkieli, joka on laajennettavissa eri tarkoituksia varten uusilla merkintäelementeillä.
- **CSV:** Comma Separated Value (CSV) tiedostomuoto, jossa arvot erotetaan toisistaan pilkuilla. Tiedostot voidaan avata taulukkolaskentaohjelmilla.

- **JSON:** JavaScript Object Notation (JSON) on kevyt ohjelmointikielistä riippumaton tekstipohjainen datan siirtoformaatti.
- **RDF:** Resource Description Framework (RDF) on linked data -paradigman standardi, jossa yksittäisiä tietoresursseja kuvaillaan niihin linkitettävien sanastojen avulla. (Tässä luvussa RDF-lyhenteellä viitataan XML-muotoisiin RDF-tiedostoihin, joiden nimeämisessä monasti käytetään .rdf-päätettä.)

XML-pohjaisia eri tarkoituksiin kehitettyjä erikoistuneita merkintäkieliä

- **RSS ja GeoRSS:** Really Simple Syndication (RSS) on XML-pohjainen tietoformaatti päivittyvien verkkosisältöjen eli syötteiden välitykseen. Mikäli syötteeseen liitetään paikkakoordinaatit, on kysessää GeoRSS-syöte.
- **ATOM:** Atom nimi viittaa kahteen toisiaan lähellä olevaan standardiin. Atom Syndication Format on XML-pohjainen merkintäkieli verkkosyötteiden (vrt. RSS) esittämiseen ja Atom Publishing Protocol (AtomPub) on yksinkertainen HTTP-protokollaan pohjautuva ohjelmointirajapinnan kuvaava standardi blogien päivittämiselle.
- **KML:** Keyhole Markup Language (KML) on XML-pohjainen merkintäkieli paikkatietojen merkitsemiseen ja niiden näyttämiseen karttapalveluissa.

Asiakirjojen esittämiseen tarkoitettuja tiedostomuotoja

- **HTML:** Hypertext Markup Language (HTML) on www:n keskeisin tiedostomuoto, joka mahdollistaa verkkosivujen rakenteen (mutta ei sisällön rakenteen) esittämisen ja sivujen linkittämisen toisiinsa yhdeksi hypertekstiverkoksi.
- **PDF:** Portable Document Format (PDF) internetissä yleinen dokumentin hyvää tulostettavuutta painottava tiedostomuoto.

Siirtoprotokollia

- **HTTP:** Hypertext Transfer Protocol (HTTP), eli hypertekstin siirtoprotokolla on yksi internetin keskeisimmistä standardeista.
- **PuSH:** PubSubHubbub (PuSH) siirtoprotokolla nopeaan tietopäivityksistä ilmoittamiseen.

Tekniikoita metatietojen sisällyttämiseen verkkosivuille

- **Mikroformaatit** ovat joukko pieniä HTML:n elementeistä muodostettuja formaatteja koneluettavan tiedon upottamiseksi www-sivuihin.
- **RDFa:** formaatti koneluettavien merkityksien upottamiseksi www-sivuihin.

Resurssien nimeäminen

- **URI:** Universal Resource Identifier (URI) on internet-resurssin tunniste

Arkkitehtuurityyli rajapintojen tekemiseen

- **REST:** Representational State Transfer (REST) on HTTP-protokollaan perustuva arkkitehtuurimalli ohjelmointirajapintojen toteuttamiseen.

5.1 Suunnittelukysymyksiä

Datan koneluettavuus ei ole suoraviivainen mekaaninen toimenpide. Seuraavassa esitämme muutamia tyypillisiä eri toteutusvaihtoehtoihin liittyviä kysymyksiä, jotka nousevat esille siinä vaiheessa, kun jokin aineisto päätetään julkaista verkkoon:

- Missä formaatissa julkaistaan?
- Mitä standardeja käytetään?
- Miten julkaistava tieto pitäisi kuvailla ja mitä metadataa tarjota?
- Pitäisikö tehdä rajapinta vai tarjota aineisto ladattavaksi tiedostomuodossa?

Avoin lähdekoodi, avoin data ja avoimet formaatit ovat tietohallinnon strategisia työkaluja, joiden avulla voidaan välttää riippuvuus yksittäisestä toimijasta. Julkisen hallinnon tietohallinnon neuvottelukunta (JUHTA) julkaisi vuonna 2009 suosituksen avoimen lähdekoodin ohjelmien käytöstä julkisessa hallinnossa (JHS 169). Siinä missä avoin lähdekoodi mahdollistaa sen, että hallinnon ohjelmistokehitys ei ole yhdestä toimittajasta riippuvaista, avoin data mahdollistaa sen, ettei hallinnolla ole tahatonta monopolia uusien palveluiden ideointiin ja kehittämiseen. Luonnollisesti avoimen ja myös suljetun datan siirrossa tarvitaan alusta- ja ohjelmistoriippumattomia siirtoformaatteja.

Standardeja tarvitaan. Avoimet standardit ovat edellytyksiä avoimille markkinoille. Vaikkei standardeja arkisesti tule ajatelleeksi, niin ilman niitä tuotteen ostajat olisivat aina riippuvaisia valmistajasta eikä jo hankittuihin tuotteisiin voitaisi hankkia lisäosia tai muuta yhteentoimivaa tuotetta. Jokainen meistä pitää itsestäänselvänä, että ostamamme energiansäästölamppu sopii aikaisemmin omistamaamme valaisimeen. Tämä ajattelumalli ei ole vielä kovin vanhaa tietotekniikka-alalla. Esimerkiksi vasta hiljattain on saatu EU:n avustamana matkapuhelinvalmistajat sopimaan siitä, miten matkapuhelimen laturi liitetään puhelimeen. Tulevaisuudessa matkapuhelimen ostaja ei enää ole riippuvainen puhelimen valmistajasta, mikäli laturi häviää ja hänen pitää ostaa tai lainata sitä (EU 2009).

Avoin standardointi ei järjestelmäntoimittajan tärkeysjärjestyksessä aina ole kovinkaan korkealla. Entuudestaan vahvoilla oleva valmistaja hyötyy siitä, että asiakkaat ovat riippuvaisia sen tuotteista esimerkiksi valmistajan standardoimattomien tiedostomuotojen käytön yleisyyden vuoksi. Kilpailutilanteen muuttuessa vahvoilla ollut toimittaja saattaa myöhemmin tehdä myönnytyksiä ja julkistaa standardin muiden käytettäväksi, mutta pitää edelleen standardin kehitystyön itsellään tai perii maksua standardin käytöstä.

5.2 Mitä on koneluettavuus

Merkintäkielenä HTML (taulukko 5.1) soveltuu hyvin asiakirjojen rakenteen kuvailuun (mikä on otsikkoa ja mikä leipätekstiä), mutta itse tietosisältöä se ei tarjoa koneluettavassa muodossa. Esimerkiksi jos kunnan verkkosivuilla luetellaan kaikkien kirjastojen osoitteet ja aukioloajat on ihmisen helppo löytää haluamansa tieto sivuilta, mutta koneellisesti on vaikea erottaa osoitteita, aukioloaikoja ja muuta sivuilta löytyvää informaatiota toisistaan.

Jos asiaa katsotaan hetken kuvitteellisen “Kunnallispalvelut kännykässä” mobiilisovellusta kehittävän ohjelmoijan näkökulmasta, joutuisi hän pahimmillaan kopioimaan aukiolotiedot käsin ohjelmaansa internet-sivulta - tiedonmurusta toiseen. Ensimmäinen askel ohjelmoijan avustamiseksi olisi tarjota osoite- ja aukiolotiedot pilkku- tai sarkainerotellussa tekstitiedostossa (esim. CSV) ladattaviksi. Tällöin ohjelmoija voisi kerralla lukea kaikki kirjastojen osoitteet ja aukioloajat ohjelmaansa koneellisesti.

Yksinkertaistetussa esimerkissämme ohjelmoijan seuraava haaste on eri lähteistä kerättyjen kirjastojen, päiväkotien ja uimahallien osoite- ja aukioloaikatietoja kuvaavien aineistojen yhdistäminen. Koneluettavuuden näkökulmasta on pystyttävä yhdistämään eri ainestojen metatiedot, jotta koneellisesti voidaan todeta, kirjastojen *osoitteiden* olevan verrannollisia *katuosoitteeseen* uimahalleista puhuttaessa. Aina tietueiden yhdistäminen (harmonisoiminen) ei kuitenkaan ole näin suoraviivaista. Hyvän esimerkin tarjoavat työttömyysluvut, jotka Tilastokeskus ja työ- ja elinkeinoministeriö laskevat eri tavalla. Näitä lukuja ei voida helposti yhdistää.

Pilkku- tai sarkaineroteltu data sopii parhaiten taulukkomuotoisen tiedon tai nimi-arvoparien julkaisemiseen. Etuna on, että lähes jokaiselta tietokoneelta löytyy taulukkolaskentaohjelma, jolla tällaiset tiedostot voidaan avata tutustuttaviksi ja muokattaviksi. Monimutkaisempien tietorakenteiden esittämiseen ainakin XML, JSON ja RDF ovat mainitsemisen arvoisia yleisen tason standardeja.

Koneiden välisessä tiedonsiirrossa ja automaattisessa prosessoinnissa verkon *lingua franca* on XML. Se ei ole valmis formaatti tiedon julkaisemiseen, vaan standardi, jolla voidaan määritellä sovelluskohtaisia merkkauskieliä tietojen esittämiseksi. Monimutkaisten tietorakenteiden julkaisemiseen voidaan käyttää jotakin olemassa olevaa XML:ään perustuvaa kieltä tai määritellä oma XML-rakenne. Eri merkkauskielien harmonisoimiseksi linked data -paradigmassa tietueet esitetään RDF-muodossa (kohta 5.6).

JSON on puolestaan XML:ää kevyempi tapa esittää ohjelmallisesti käsiteltävää dataa mahdollisimman helposti siirrettävässä muodossa. JSON-syntaksin avulla voidaan esittää sekä yksinkertaisia nimi-arvopareja että monimutkaisia tietorakenteita siten, että niitä on helppo käsitellä paljon käytettyjen internet-ohjelmointikielten avulla.

5.2.1 Rajapinnat ja esitysmuodot datan mukaisesti

Kun aineistojen avaamisesta on päätetty, on usein ensimmäisiä kysymyksiä, missä formaatissa dataa pitäisi jakaa. Tiedon käyttäjille tärkeintä on, että data on saatavilla ohjelmallisesti käsiteltävässä muodossa ja esitysmuoto on avoin, eli että sen käsittely ei edellytä minkään tietyn toimittajan ohjelmistoja. Esitysmuodon avoimuus ja koneluettavuus ovat minimivaatimukset, jotta dataa päästään koneellisesti käsittelemään. Esitysmuodot liittyvät datan käytön helppouteen (tai vaivalloisuuteen) myös siten, että jokainen muunnos muodosta toiseen vaatii työtä; jotkut muunnokset ovat helpompia kuin toiset.

Pilkku- tai sarkaineroteltu data sopii parhaiten taulukkomuotoisen tiedon tai nimi-arvoparien julkaisemiseen. Etuna on, että lähes jokaiselta tietokoneelta löytyy taulukkolaskentaohjelma, jolla tällaiset tiedostot voidaan avata tutustuttaviksi ja muokattaviksi. Monimutkaisempien tietorakenteiden esittämiseen ainakin XML, JSON ja RDF ovat mainitsemisen arvoisia yleisen tason standardeja.

Koneiden välisessä tiedonsiirrossa ja automaattisessa prosessoinnissa verkon ”lingua franca” on XML. Se ei ole valmis formaatti tiedon julkaisemiseen, vaan standardi, jolla voidaan määritellä sovelluskohtaisia merkkauskieliä tietojen esittämiseksi. Monimutkaisten tietorakenteiden julkaisemiseen voidaan käyttää jotakin olemassa olevaa XML:ään perustuvaa kieltä (esim. paikkatietojen esittämiseen käytetty KML) tai määritellä oma XML-rakenne. Eri merkkauskielien harmonisoimiseksi linked data -paradigmassa tietueet esitetään RDF-muodossa (kohta 5.6).

JSON on puolestaan XML:ää kevyempi tapa esittää ohjelmallisesti käsiteltävää dataa mahdollisimman helposti siirrettävässä muodossa. JSON-syntaksin avulla voidaan esittää sekä yksinkertaisia nimi-arvopareja että monimutkaisia tietorakenteita siten, että niitä on helppo käsitellä paljon käytettyjen internet-ohjelmointikielten avulla.

Ensivaiheessa datan tarjoajan kannattaa saattaa data päivänvaloon itselleen helpoimmalla tavalla. Myöhemmin datan käytettävyyttä voidaan lisätä tekemällä muunnoksia muihin esitysmuotoihin. Nyrkkisääntönä voidaan pitää, että muunnokset yleisesti käytettyihin esitysmuotoihin kannattaa tehdä datan tarjoajan päässä sen sijaan, että jokaisen datan hyödyntäjän pitäisi tehdä sama muunnos erikseen. Sama aineisto voidaan tarjota esimerkiksi XML-, JSON- ja RDF-muodossa (kohta 5.5).

Kun data on koneluettavasti saatavilla, se on dokumentoitava, jolloin voidaan kertoa mitä mikäkin tieto tarkoittaa. Esimerkiksi pilkku- tai sarkaineroteltuna voidaan koneluettavasti esittää taulukkomuotoista tietoa, mutta se ei ota kantaa siihen, mitä sarakkeita taulukossa on. Ilman sarakeotsikoita moni taulukko olisi melko hankalasti käsitettävissä. Siksi koneellisesti käsiteltävän muodon ohella on tärkeää tarjota dokumentaatio siitä, mitä rakenne pitää sisällään. Esimerkkitapauksessa osoitteista ja aukioloajoista voi sisällöstä päätellä merkityksen, mutta yleensä tapaukset ovat monimutkaisempia ja selkeä tiedon käyttöopas on tarpeen.

5.2.2 Koneluettavat lisenssit

Samalla kun huolehditaan datan tarjoamisesta koneluettavassa muodossa, on hyvä tarjota myös aineiston käyttölisenssi samalla tavalla. Lisenssien koneluettavuus helpottaa yhdistelmäpalveluiden kehittäjien mahdollisuuksia seurata ehtojen noudattamista, sillä käyttöehdot voidaan huomioida palvelun toiminnoissa.

Koneluettavuus parantaa myös aineistojen löydettävyyttä. Esimerkiksi Flickr-kuvapalvelusta on mahdollista hakea hakusanaa vastaavia kuvia, joiden lisenssiehdot sallivat niiden uudelleenkäytön vaikkapa omassa powerpoint-esityksessä. Creative Commons -lisensseillä (Liite 4) on ollut mahdollista ilmaista käyttöoikeudet koneluettavasti jo vuosikausia. Vaikka aineisto olisi julkishyödyke, on sekin syytä ilmaista eksplisiittisesti. Tämä tarve vain korostuu koneluettavuuden myötä. Creative Commons Waiver -lisenssi (CC0) sopii tähän tarkoitukseen. Myös PSI-direktiivissä suositellaan lisenssien tarjoamista koneluettavassa muodossa.

5.3 Web-arkkitehtuuri

Termejä www ja internet käytetään usein erheellisesti toistensa synonyymeinä. Www on kuitenkin vain yksi monista internetin palvelumuodoista. Muita

palvelumuotoja ovat esimerkiksi sähköposti, keskusteluryhmät tai tiedoston siirto. Internetille luotiin pohjaa jo 1960-luvulla Yhdysvalloissa, ja nykyisin se on maailmanlaajuinen eri tietoverkkoja yhdistävä laitteisto- ja ohjelmistoinfrastruktuuri, jonka avulla tietokoneet voivat olla yhteydessä toisiinsa. Tim Berners-Leen 1990-luvulla kehittämä www on puolestaan internet-verkossa toimiva hajautettu linkkeihin perustuva hypertekstijärjestelmä. Kun tässä oppaassa puhutaan tietovarantojen julkaisemisesta avoimesti verkkoon, tarkoitetaan sillä sitä, että ne tuodaan osaksi www:tä ja web-arkkitehtuuria.

W3C-standardointiorganisaatio edistää www:n kehittymistä rakentaen yhteisiä sopimuksia edistämään webin eri osien yhteensopivuutta. W3C sai vuonna 2004 valmiiksi suosituksen, joka kirjaa kokonaisuutena toimivan web-arkkitehtuurin keskeiset osat ja suunnitteluperiaatteet (W3C 2004). Valitettavasti sen suosituksia ei aina tunneta tai noudateta esimerkiksi ohjelmallisesti luotujen URI-tunnuksien muodostamisen osalta. Suositus tarjoaa hyvän kehikon datan tuomiselle verkkoon.

Web-arkkitehtuuri mahdollistaa asteittaisen kehityksen hajautetussa verkkoympäristössä. Se on osoittanut muuntautumiskykyä ja yhden orgaanisesti kasvavan linkittyneen informaatioavaruuden voiman. Myöhemmin tässä luvussa käsiteltävä linkitetty data esittelee, miten dataa voidaan julkaista intenetissä kytkeytyen osaksi maailmanlaajuista www-informaatioavaruutta. Samoin rajapintojen yhteydessä esiteltävät REST-tyylin palvelurajapinnat perustuvat web-arkkitehtuuriin ja soveltuvat siksi hyvin datan julkaisemiseen avoimessa verkossa.

Verkossa toimii voimakas pienimmän yhteisen nimittäjän voima. Monikäyttöiset, yksinkertaiset ja helposti omaksuttavat ratkaisut, joilla voidaan kattaa laajoja sovellusalueita, saattavat levitä yllättävänkin laajaan käyttöön. Aikaa ja vaivaa säästyy, kun teknologia-komponentteja, osaamista ja käytäntöjä voidaan soveltaa ongelmien ratkaisemiseen. Näin on käynyt www:lle, kun siitä on tullut yleinen väline erilaisten palveluiden toteuttamiseen. Käytännössä melkein koko päivittäin käyttämämme internet perustuu muutamaan erittäin laajasti omaksuttuun standardiin (TCP/IP, HTTP, HTML, CSS, JavaScript jne.). Nämä standardit ovat tuttuja kehittäjille, ylläpitäjille, ohjelmistoarkkitehdeille ja web-arkkitehtuuriin perustuvien järjestelmien toteuttamiseen löytyy paljon valmiita avoimen lähdekoodin ohjelmistoja ja edistyksellisiä työkaluja.

Yksi tärkeimmistä työkaluista on päivittäin käyttämämme selain (kuva 5.1). Alunperin hypertekstin selaamiseen tarkoitettujen selainohjelmien ominaisuudet ovat laajentuneet ja niitä käytetään www:n ohella käyttöliittyminä moniin muihinkin internetin palvelumuotoihin, kuten edellä mainittuihin sähköpostiin ja tiedonsiirtoon, sekä pikaviestintään sosiaalisen median palveluissa. Www-termikin alkaa hävitä puhekielestä.

Web-arkkitehtuurin keskeisin käsite on globaali informaatioavaruus, joka koostuu yksiselitteisesti tunnistettavista toisiinsa linkitetyistä resursseista. Resurssi voi esimerkiksi olla dokumentti tai verkkoon kytketty tietokoneohjelma, jolla

on yksikäsitteinen tunniste. Kaikille verkon käyttäjille tuttuja URI-tunnisteita ovat internetosoitteet, kuten http://www.suomi.fi. Web arkkitehtuurin piiriin tuotavien asioiden nimeämiseen suositellaan käytettäväksi HTTP-URI:a muiden merkintätapojen sijaan.

Sama resurssi - vaikkapa tieto eduskunnan täysistunnon äänestystuloksista - voidaan esittää useissa esitysmuodoissa, kuten HTML, CSV, XML tai RDF (ks 5.4). Web-arkkitehtuurin ytimessä on hypertekstin siirtoprotokolla HTTP (Hypertext Transfer Protocol), joka määrittelee mahdolliset toiminnot (GET, PUT, POST ja DELETE) selaimien ja www-palvelimien väliseen vuorovaikutukseen. Tämä perusrakenne on osoittautunut toimivaksi sekä verkossa surffaavien ihmisten että tietokoneohjelmien ohjelmoimisen näkökulmasta.

Kuva 5.1: Hypertekstiä luetaan selaimella, joka hakee verkkosivuiksi kutsuttuja html- dokumentteja verkkopalvelimilta ja esittää ne ihmisille luettavassa muodossa.

5.4 Sisällön esitysmuotoja

Internetsovelluksissa käytetään IANA:n (Internet Assigned Numbers Authority) standardoimia sisällön esitysmuotoja eli niin sanottuja MIME-mediatyyppejä. Esitysmuotoihin on listattu yleisimpien tietokoneohjelmien tiedostomuodot sekä vain internetissä esiintyviä esitysmuotoja. Avoimien esitysmuotojen käsittelyyn on olemassa paljon ohjelmistokirjastoja ja työkaluja eri ohjelmointikielille. Jotkin valmistajakohtaiset esitysmuodot ovat myös yleisiä internetissä, kuten Microsoft Excel-tiedostomuodot. Näiden käyttö on sidotumpaa valmistajan omiin työkaluihin.

5.4.1 Ilmoitukset muutoksista syötteiden avulla

Usein päivittyvät tiedot voidaan esittää käyttäjälle tietovirtana, josta näkee ajantasaisesti tuoreimmat muutokset. Paras tapa ilmoittaa muutoksista on julkaista syöte. Syötteen tilaajat saavat automaattisesti tiedon muutoksesta omaan järjestelmäänsä (esim. syötteiden lukijaan, selaimeen tai vaikka sähköpostiohjelmaan). Syötteissä julkaistavia tietoja on myös melko helppo hyödyntää ohjelmallisesti, joten joissakin tapauksissa ne kelpaavat sellaisenaan datan jakamiseen. Syötteiden ilmaisuvoima koneluettavien merkityksien ilmaisemiseen on kuitenkin heikko. Muutoksista ilmoittamisen ohella syötteet ovat hyvä tapa välittää otsikoita, lyhyehköjä tekstejä ja asiaan liittyviä linkkejä, kuten Hovioikeuksien ratkaisujen tapauksessa on tehty (http://www.oikeus.fi/rss/ho/hovioikeuksienratkaisut.rss - löytyy suomi.fi/dataopas).

Syötteiden kuvaamiseen on olemassa kaksi yleisesti käytettyä esitysmuotoa: RSS ja Atom. Teknisesti RSS ja ATOM ovat XML-merkintäkieliä. Jakamalla tuottamansa tiedon syötetiedostomuodossa sisällöntuottaja mahdollistaa

tiedon uudelleenjulkaisemisen muissa verkkopalveluissa ja tiedon säännöllisen seuraamisen syötteenlukuohjelmilla. RSS-muotoa käytetään nykyisin erittäin laajasti mm. uutisten, blogien ja muun ajankohtaisen sisällön, kuten tiedotteiden, välittämiseen.

5.4.2 Ajantasainen verkko

Uusien teknologioiden ansiosta dataa voidaan tuottaa ja seurata maailmanlaajuisesti ja ajantasaisesti. Ajantasaisen datan jakelu ja seuraaminen (real-time web) on nopeasti lisännyt merkitystään internetin käytössä. Miten sitten julkishallinnon data liittyy reaaliaikaiseen internetiin? Terveyskeskusten jonotustilanne, junien pysyminen aikataulussa ja säätiedot ovat esimerkkejä sovelluskohteista, joissa ajantasainen data helpottaisi ihmisten elämää. Viivästyneenä saman datan käyttömahdollisuudet ovat rajallisempia. Suurin hyöty datasta saadaan julkaisemalla se sekä ajantasaisesti että arkistoimalla se myöhempää käyttöä varten. Tällöin dataa voidaan jälkikäteen käyttää esimerkiksi toiminnan kehittäseen.

Ongelmana ajantasaisuuteen pyrkimisessä on se, että internetiä ei ole alunperin suunniteltu ajantasaisuutta varten. Tähän asti verkkosovellukset toimivat uuden informaation saamiseksi kyselyperiaatteella lähettäen julkaisijalle jatkuvasti *onko mitään uutta* -kyselyitä, vaikka 99 % tapauksista vastaus on *ei*. Ajantasaisten paveluiden tekemiseen on vasta hiljattain yleistynyt protokollia, joilla tämä toimintalogiikka on käännetty toisinpäin eli tiedon julkaisijataho lähettää halukkaille tilaajille uuden informaation silloin, kun sitä on. Tällä hetkellä voimakkaimmin suosiotaan kasvattava protokolla ajantasaisessa internetissä on PubSubHubbub (PuSH), jonka avulla uusi sisältö lähetetään suoraan tilaajalle.

5.4.3 Paikkatieto uudelleenkäytettävässä muodossa

Moni tieto liittyy tiettyyn fyysiseen sijaintiin, kuten paikkaan, alueeseen tai osoitteeseen. Paikkatieto laajasti käsitettynä kattaakin merkittävän osan julkishallinnon tuottamasta informaatiosta, sillä sijaintitieto yleensä koordinaattien, aluetunnusten tai osoitteiden muodossa liittyy moniin hallinnon ylläpitämiin rekistereihin. Paikkatiedon erityispiirteenä on, että sen hallinnointi on ollut perinteisesti varsin ammattimaista ja verrattain teknisorientoituneen henkilökunnan käsissä, minkä ansiosta sen uudelleenkäytettävyyden ja jakamisen eteen on tehty paljon enemmän töitä kuin esimerkiksi tekstimuotoisen asiakirjatiedon.

Jos tieto sijainnista on koneellisesti käsiteltävässä muodossa, niin tietoja voidaan esittää havainnollisesti kartalla ja eri lähteiden tietoja voidaan yhdistellä toisiinsa niiden sijainnin perusteella. Nykyisin karttapalveluiden teko on ohjelmoijille melko yksinkertaista olemassa olevien rakenteiden, kuten Open Streetmap tai Google Maps, avulla. Paikkatiedon uudelleenkäyttö on

yleistyneiden karttasovellusten muodossa helposti ymmärrettävää, mikä on vilkastuttanut sen ympärillä käytyä keskustelua.

Taulukko 5.2. Esimerkkejä tiedoista, joihin tyypillisesti liittyy sijainti
Päätöksenteko kaavamuutosehdotukset, tiettyä aluetta tai osoitetta koskevat esitykset ja päätökset,... Palvelut kaupalliset ja julkiset palvelut; kirjastot, sairaalat ja liikuntapaikat aukioloaikoineen; kaupat, ravintolat, elämyspalvelut,... Liikenne ruuhkatiedot, tietyöt, julkisen liikenteen pysäkit aikatauluineen sekä poikkeusliikennetiedotukset vaikutusalueineen, ajoneuvojen ajantasaiset sijainnit,... Säätiedot tiesää, sääennusteet, ... Mediasisältö valokuvat, videot, johonkin paikkaan liittyvät tarinat,... Teknisesti sijaintitiedon lisääminen esimerkiksi RSS-syötteenä julkaistaviin tietoihin on melko yksinkertaista. Sijaintimerkinnät sisältävää syötettä kutsutaan GeoRSS-syötteeksi. Google Maps -ohjelmointirajapinnassa on valmiit toiminnot GeoRSS-syötteen tietojen esittämiseen kartalla. Vielä monipuolisemmin sijaintitietoa voidaan julkaista KML Network Link -toiminnallisuuden avulla.

Google Maps ja Google Earth osaavat hyödyntää KML-syötteitä niin, että käyttäjälle haetaan tietoja aina näkyvissä olevalta alueelta ja tietoja päivitetään näkymän muuttuessa. Tästä syystä syötteenä kannattaa julkaista myös sellaisia tietoja, jotka eivät itsestään päivity, mutta joista tulee käyttäjälle merkityksellisiä, kun käyttäjä tarkastelee tiettyä aluetta. Samaa GeoRSS- tai KML-syötettä voidaan hyödyntää useassa karttapalvelussa (tai muussa sijaintitietoa hyödyntävässä palvelussa). Toisaalta yhteen palveluun voidaan liittää GeoRSS- ja KML-syötteitä useista lähteistä.

5.4.4 Asiakirjojen julkaiseminen

Suomen kielen perussanakirjan mukaan asiakirja on *"määrätarkoitukseen käytettävä kirjallinen esitys"*. Asiakirjoja voidaan toki säilyttää myös digitaalisessa muodossa, mutta ajatuksellisesti sana viittaa usein kiinteisiin artifakteihin, kuten sopimuspaperit tai passi. Vähintäänkin vaaditaan, että digitaalinen asiakirja pitää voida tulostaa. Tyyppiesimerkki nykyään täysin tarpeettomasta asiakirjasta on lentolippu. Matkatoimisto lähettää sen sähköpostitse ja matkustaja tulostaa varmuuden vuoksi. Monissa tapauksissa olisikin perustellumpaa julkaista ja välittää asiakirjojen sisältämä informaatio sellaisenaan.

Kaupunginvaltuuston kokousten asialista on tyyppiesimerkki asiakirjoista. Nykyisin kaupungin hallintokoneistoissa erilaiset päätösasiat etenevät aloitteista kokouksien kautta lausunnoille, lautakuntiin ja taas uusiin kokouksiin. Päätösasioita hallinnoidaan usein erillisillä asianhallintajärjestelmillä, joista sitten yksittäisen kokouksen asialistat julkaistaan ennen kokousta internetiin. Yksittäisen asian etenemisestä kiinnostunut joutuu helposti käymään läpi ison pinon pitkiä asialista- ja päätöspöytäkirjadokumentteja löytääkseen etsimänsä. Mikäli hänellä olisi sen sijaan mahdollisuus tehdä hakuja asian-

hallintajärjestelmän sisäiseen tietokantaan, saisi hän yhdellä tai muutamalla haulla koottua itselleen merkityksellisen informaation.

5.5 Rajapinnat, sovellukset ja palvelut

Mitä tarkoitetaan, kun puhutaan avoimista rajapinnoista? Tietokoneohjelmien käyttöliittymä on ihmiskäyttäjän ja ohjelman välinen rajapinta, jonka kautta kommunikaatio tapahtuu ja informaatio liikkuu. Vastaavasti ohjelmien välillä on rajapintoja, joiden kautta ohjelmat tai ohjelmiston osat kommunikoivat keskenään. Usein rajapinnat on tarkoitettu vain järjestelmän sisäiseen käyttöön ja mahdollisiin ennalta sovittuihin tietotekniikkajärjestelmien integrointeihin. Nykyisin on kuitenkin erittäin yleistä toteuttaa esimerkiksi verkkopalveluun avoin ns. web API (application program interface) eli verkko-ohjelmointirajapinta, jota on mahdollista käyttää internetin välityksellä.

Tyypillisesti web API on toteutettu järjestelmään, joka voi tarjota erilaisia valmiita palveluita muille ns. sovellusohjelmille. Tämä helpottaa ohjelmointityötä, kun kaikkea ei tarvitse tehdä moneen kertaan. Ohjelmoijien suosima Googlen karttarajapinta tarjoaa esimerkiksi kätevän geokoodauspalvelun, joka muuntaa tekstimuotoiseen osoitteen (ja paljon muutakin) vastaavaksi maantieteelliseksi koordinaatiksi. Tällaisen ohjelmoiminen itse olisi työlästä, vaikka kaikki tarvittava data olisikin käytössä.

Rajapinnan kautta tarjottava palvelu voi olla yksinkertaisimmillaan sellainen, että lähettämällä sopivan pyynnön palvelu vastaa palauttamalla tiedoston. Näin yksinkertaista palvelua ei mielletä välttämättä palveluksi, joten puhutaan pelkistä rajapinnoista, joiden kautta saa dataa. (Tarkoin termein tällaisia kutsutaan tietopalvelurajapinnoiksi.)

Palvelurajapinnat on yksi tapa tarjota dataa koneluettavassa muodossa tietoteknisten järjestelmien käytettäväksi. Esimerkkejä palveluista, jotka tarjoavat alunperin julkishallinnon tuottamaa dataa rajapintojen kautta, ovat Google Transit API sekä brittiläisen Guardian-lehden World Government Data API. Google Transit API tarjoaa julkisen liikenteen reittitietoja maailmanlaajuisesti ja Guardianin rajapinnan kautta käsiksi tällä hetkellä Iso-Britannian, USA:n, Australian, Uuden Seelannin ja Lontoon julkisen datan katalogien sisältöön.

Ohjelmille rajapintojen kautta tarjottavia palveluita ei pidä sekoittaa HTML:ään perustuviin ihmisten luettavaksi ja käytettäväksi tarkoitettuihin verkkopalveluihin. Esimerkiksi Helsingin seudun liikennelaitoksen reittiopas (ennen YTV:n reittiopas) on yksi Suomen suosituimmista verkkopalveluista, mutta reittiopasjärjestelmään on olemassa myös ohjelmointirajapinta, jonka kautta voidaan ohjelmallisesti hakea muun muassa aikataulu- ja pysäkkitietoja. Tällöin reittiopasjärjestelmä tarjoaa datan rajapinnan kautta palveluna muille sovellusohjelmille, kuten vaikkapa iPhone-kännykkäsovellukseen, jonka kautta reittejä voi suunnitella. Liitteessä 5 on käsitelty tarkemmin yleinen

palvelukeskeinen arkkitehtuuri (SOA Service Oriented Architecture) ja ns. REST -tyylin rajapinnat, jotka ovat suosittuja avoimen datan jakamiseen.

5.6 Linkitetty data

Toisiinsa linkitettyihin resursseihin perustuva web-arkkitehtuuri on maailmanlaajuinen tietoarkkitehtuuri, johon avoimesti dataa julkaisevien tahojen tulisi datavarantonsa liittää, jotta ne olisivat tehokkaimmin kaikkien käytettävissä. Kun aikanaan ihmiset alkoivat spontaanisti solmia yhteen HTML-sivuja linkkien ja URL-osoitteiden avulla, muodostui internet sellaiseksi kuin me sen nykyään arkisesti tunnemme, ts. verkoksi toisiinsa linkitettyjä dokumentteja. Vastaavalla tavalla on mahdollista ja nykyisin jo yleistä linkittää yhteen HTML-sivujen ohella myös dataa. Kehitys on synnyttämässä uudenlaista internetiä, verkkoa toisiinsa linkitettyä dataa. Termillä linkitetty data (Linked Data) viitataan yleensä www:n kehittäjän Tim Berners–Leen vuonna 2006 julkaisemaan neljään periaatteeseen (Taulukko 5.3).

Taulukko 5.3: Linkitetyn datan neljä periaatetta (Berners-Lee 2006) Teknisemmin: Tarkoituksena:

1. Käytä URI:ja (Universal Resource Identificator) niminä asioille
    1. muodostaa asiasta käsite, josta voidaan "puhua" (johon voidaan viitata muualta)
2. Käytä HTTP URI:ja niin, että ihmiset voivat etsiä ja tarkistaa niitä nimiä
    2. tarjota käsitteeseen liittyvää tietoa sieltä, mistä sitä luontevasti etsisi
3. Kun joku tarkistaa antamasi URI:n, tarjoa hyödyllistä informaatiota käyttämällä standardeja (RDF, SPARQL)
    3. tehdä helpoksi lisätiedon löytäminen nimetyistä objekteista ja resursseista
4. Sisällytä linkkejä muihin URI:hin, jotta he voivat löytää lisää asioita
    4. muodostaa käsitteiden välisiä suhteita – luoda erillisten tiedonpalasten sijaan koko ajan laajeneva verkko

Riippumatta siitä, missä formaatissa tieto julkaistaan ja missä formaatissa se alunperin on, RDF on käyttökelpoinen malli yhdistämään tietovarantoja keskenään internetin välityksellä. Teknologiana RDF mahdollistaa asioiden ja konseptien helpon linkittämisen toisiinsa sekä toisistaan riippumattomien ja erillään suunniteltujen järjestelmien myöhemmän yhdistämisen toisiinsa. Koska RDF mahdollistaa saman datan kuvaamisen erilaisilla sanastoilla,

voidaan yhteentoimivuutta harmonisoinnin ja sanastotyön avulla lisätä siellä, missä se kannattaa kustannustehokkaasti tehdä.

Avoimen datan julkaisun yhteydessä ei ole tarpeen etsiä yhteisiä standardeja ja tapoja kuvailla kaikkea dataa, kuten vaikkapa koulua koskevaa dataa. Yksilöillä ja organisaatioilla on hyvin poikkeavia näkökulmia. Kaupungin tilapalvelun rekistereistä koulu löytynee vuokralaisena, opetusministeriöllä on paljon kouluihin liittyvää dataa ja yksittäisellä koululla itsellään on dataa muun muassa aukioloajoista. Ei ole syytä olettaa, että kaikki toimijat alkaisivat käyttää yhtä standardoitua sanastoa kuvaamaan koulua. Jos tällainen sanasto saataisiinkin sovittua, ei se kompromissina välttämättä enää palvelisi kunnolla mitään tahoa.

Yleisesti hyväksytyt kuvailusanastot lisäävät merkittävästi datavarantojen yhteiskäyttöisyyttä ja siksi niiden vapaaehtoista syntymistä kannattaa tukea. Internetissä toimiva pienimmän yhteisen nimittäjän voima (kohta 6.3) ei ole millään tavalla estänyt yksilöitä ja yhteisöjä rakentamasta yhteisen pohjan päälle heitä tukevia yksityiskohtaisempia käytäntöjä. Esimerkiksi www-sivujen merkintäkieli HTML on hyvin laajasti hyväksytty standardi, mutta sen päälle on vapaaehtoisesti ja tarvelähtöisesti kehitetty suppeampia käytäntöjä, kuten mikroformaatit, joilla vaikkapa osoitetiedot voidaan merkitä HTML:n sekaan koneymmärrettävässä muodossa. Jos tällaisia suppeampia standardeja käytetään, on tärkeää toimia niin, että ne eivät millään tavalla rajoita muun sivun käyttämistä. Sivusto näkyy, vaikka selain ei mikroformaattia ymmärtäisikään eikä yhden mikroformaatin käyttäminen yhdessä kohtaa sivustoa edellytä, että koko sivustolla käytettäisiin kaikkia mikroformaatteja.

Sanastotyössä RDF tarjoaa tasapainoa helppouden ja standardoinnin hyötyjen välillä. Dataa julkaisevilla yksilöillä tai organisaatioilla on täysi vapaus valita, miten ne kuvailevat informaatiotaan, mutta samalla niillä on mahdollisuus jakaa omia sanastojaan ja uudelleenkäyttää osia toisten luomista sanastoista sekä luoda kuvailutietoon sopiva yhdistelmä yleisiä standardeja ja omia näkemyksiään. Toinen RDF-julkaisua puoltava tekijä on se, että siihen on jälkikäteen helppo yhdistää sanastoja.

Linkitetyn datan verkko on suunniteltu kasvamaan ja kehittymään orgaanisesti samalla tavalla kuin linkitettyjen dokumenttien verkko on kasvanut. Se kasvaa, kun ihmiset ja organisaatiot toisiltaan kysymättä spontaanisti lisäävät omia resurssejaan verkkoon ja linkittävät niitä toisiinsa. Se kasvaa, kun toimijat luovat omia sanastojaan kuvailemaan asioita. Kehitykseen kuuluu itseorganisoituvuus ja sekavuus; linkkejä rikkoutuu, uusia syntyy, sanastoja yhdistyy ja jakautuu.

> **Case:** Suomalaista semanttisen webin osaamista – TerveSuomi ja Kirjasampo
>
> Semanttisessa webissä kuvaillaan tiedon keskinäisiä suhteita tietokoneen ymmärtämässä muodossa, jotta www-sisältö olisi

helpommin koneellisesti käsiteltävää. Semanttinen web oli kaikkien huulilla muutama vuosi sitten. Vähään aikaan siitä ei ole puhuttu niin paljon, mutta se on edelleen internetin tutkituimpia ja nopeimmin kehittyviä alueita. Sen käytännön merkitys kaiken kohun jälkeen on kasvamassa vauhdilla ja nyt keskitytään toimiviin sovelluksiin. Kansainvälisesti Suomi on yksi semanttisen webin edelläkävijämaita. Tietovarantojen julkaisemiseen liittyen keskeistä semanttisen webin ajattelutavassa on linkitetty data.

TerveSuomi.fi on julkishallinnon ja järjestöjen yhteinen terveystiedon portaali, joka perustuu semanttisen webin tekniikoille ja mukana olevien toimijoiden tarjoamaan avoimeen dataan. Kun sisällöt varustetaan yhtenäisillä metatiedoilla, koneet osaavat yhdistää toisiinsa kuuluvia asioita yhteen. Näin palvelua käyttävä saa tarkempaa täsmätietoa sekalaisten hakutulosten sijaan.

Lähiaikoina valmistuva Kirjasampo-kaunokirjallisuussivusto on ensimmäinen valtakunnallinen kirjastoverkkopalvelu, joka hyödyntää semanttista webiä. Kirjasampo sisältää metadataa kaunokirjallisuudesta (myös vanhasta ja aiemmin vähälle sisällönkuvailulle jääneestä kirjallisuudesta). Rajapinnoiltaan siitä on tarkoitus tulla avoin.