Paikkatiedon käyttö web-dokumenteissa

Ilkka Vänskä

7.12.2004

Joensuun yliopisto
Tietojenkäsittelytieteen laitos
Pro gradu -tutkielma

TIIVISTELMÄ

World Wide Web sisältää valtavan määrän informaatiota, jolla on maantieteellinen luonne. Paikkatieto voi olla dokumentin kirjoittajan tietoisesti lisäämää eksaktia paikkatietoa tai tekstistä muilla keinoilla löydettävissä olevaa epäsuoraa tai pääteltävää paikkatietoa. Web-dokumenttien sisältämä paikkatieto voidaan eristää ja sitä voidaan käyttää hyväksi toteutettaessa uusia sovelluksia. Eräs esimerkki sovelluksesta, joka käyttää web-dokumenteista löytynyttä paikkatietoa, on paikallishakukone. Koko verkon sisältöä käyttäviä internet-hakukoneita ei toistaiseksi ole avoimessa käytössä, vaan haut perustuvat paikkatietorekistereihin.

Tutkimastani 24000:sta suomalaisesta verkkosivustosta 35% sisälsi epäsuoraa paikkatietoa. Eksaktia paikkatietoa löytyi häviävän vähän. Löytyneen paikkatiedon määrään vaikuttavat muun muassa käytetty sovellus, tutkittujen sivustojen kaupallinen luonne sekä verkon läpikäyntialgoritmi. Löytyneen paikkatiedon määrän perusteella arvioin, että paikallishakukoneen toteuttaminen olisi järkevää ja mahdollista.

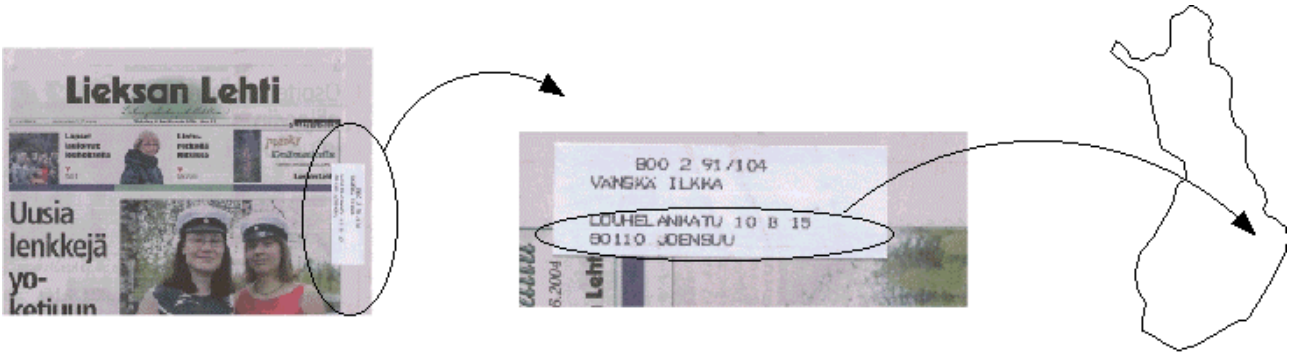
Avainsanat: paikkatieto, eksakti paikkatieto, epäsuora paikkatieto, pääteltävä paikkatieto, paikkatietorekisterit, paikallishaku, hakukoneet, paikkatiedon eristäminen.

SISÄLLYSLUETTELO

1 JOHDANTO

Paikkatietojärjestelmä (*Geographic Information System*, *GIS*) on tietojärjestelmä, joka mahdollistaa sijaintitiedon hallinnan ja prosessoinnin osana muuta järjestelmää [24]. Paikkatietojärjestelmän avulla on mahdollista toteuttaa sijaintiin perustuvia hakuja, analyysejä ja visualisointeja.

Luonnollisimmillaan paikkatiedon hyväksikäyttäminen on *mobiileissa* ympäristöissä, joissa käyttäjän sijainti ja informaation tarve voivat muuttua. Paikkatietoon voi kuitenkin törmätä hyvinkin monenlaisissa ja arkisissa tilanteissa ilman, että kyseessä on varsinainen koordinaattitasolle menevä paikantaminen. Kuvassa 1 tällaista paikkatietoa edustaa muutaman päivän vanhan sanomalehden osoitelappu, josta voi helposti ja melko luotettavasti päätellä likimäärin missä maailmankolkassa kyseinen lehti on ilmestymisensä aikoihin ollut.



Kuva 1. Lehden paikantaminen osoitelapun avulla

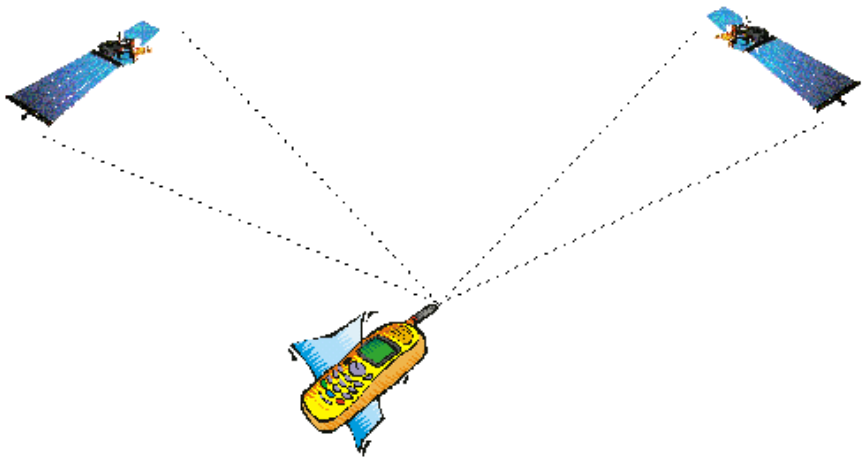
Mobiilien päätelaitteiden nopea yleistyminen on eräs merkittävimmistä syistä siihen, miksi paikkatietojärjestelmien uskotaan kehittyvän ja lisääntyvän voimakkaasti lähitulevaisuudessa. Mobiileina päätelaitteina ymmärretään esimerkiksi kämmentietokoneet, matkapuhelimet ja internet-yhteydellä varustetut kannettavat tietokoneet.

Matkapuhelinten määrä on lisääntynyt merkittävästi viimeisen reilun kymmenen vuoden aikana. Kun vuonna 1990 vain seitsemässä prosentissa suomalaisista kotitalouksista oli matkapuhelin, oli vuonna 2001 vastaava luku jo 90 prosenttia [25]. Tekniikan kehittyessä matkapuhelinta voi käyttää myös muuhunkin kuin puhumiseen. Ensimmäinen ja nykyisinkin yleisin matkapuhelimen lisäominaisuus ovat *tekstiviestit* (*Short Message Service, SMS*). Tekstiviestien käyttömahdollisuus

on kuitenkin rajallinen, eikä tekstiviestien varaan voi rakentaa kovinkaan kehittynyttä paikkatietojärjestelmää.

Uudemmista tekniikoista *General Packet Radio Service* (*GPRS*) mahdollistaa nopeamman tiedonvälittämisen matkapuhelimen avulla. Yhdessä *WAP-selaimen* (*Wireless Application Protocol*) kanssa internetin jonkinasteinen käyttäminen on mahdollista useiden matkapuhelinvalmistajien malleilla. GPRS on kuitenkin liian hidas grafiikan siirtämiseen sovelluksen käytettäväksi. Tulevat kolmannen sukupolven (3G) ja neljännen sukupolven (4G) verkot vähentänevät näitä ongelmia nopeammilla tiedonsiirtomenetelmillä.

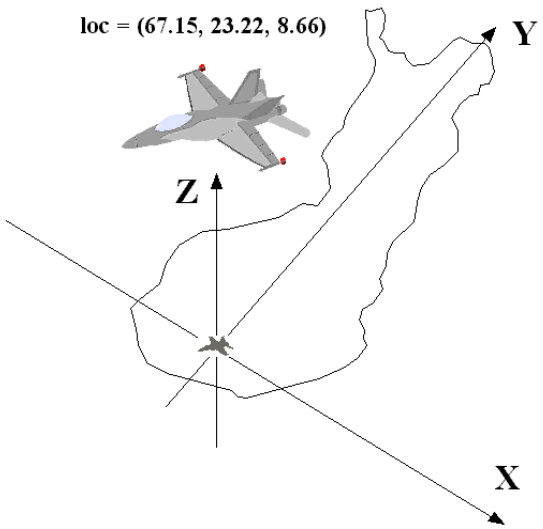
Nykytekniikka tarjoaa useampiakin mahdollisuuksia käyttäjän sijainnin määrittämiseen. Tavallinen GSM-puhelin voidaan paikantaa muutaman sadan metrin tarkkuudella [9], mikä riittää jo useisiin tarpeisiin. Yhdysvaltain sotilaskäyttöön suunniteltu, nykyisin maailmanlaajuisessa siviilikäytössä oleva satelliittipaikannusjärjestelmä *Global Positioning System* (*GPS*) mahdollistaa käyttäjän paikantamisen kymmenen metrin tarkkuudella missä vain maapallon pinnalla [9]. Suotuisissa olosuhteissa GPS-paikannus onnistuu jopa muutaman metrin tarkkuudella. Kuvassa 2 paikannukseen osallistuu kaksi satelliittia.



Kuva 2. Puhelimen paikantaminen satelliittien avulla (GPS).

Kun käyttäjä on paikannettu, puhutaan arkisesti, että hänen paikkatietonsa tiedetään. Asia on kuitenkin hieman laajempi, sillä paikkatietojärjestelmiä käsiteltäessä p*aikkatiedolla* ymmärretään kohteen sijaintitietoa ja ominaisuustietoja. *Sijaintitieto* kertoo kohteen sijainnin joko koordinaatein, osoitteen tai muun vastaavan yksikäsitteisesti kohdistettavissa olevan paikannustiedon avulla [24].

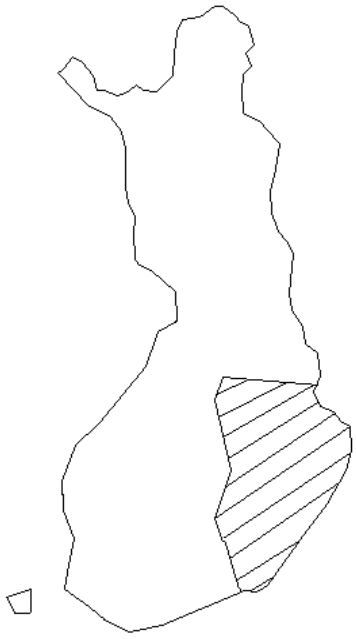
Tyypillisimmillään sijaintitieto esitetään pituus- ja leveysasteissa joko (29°45'0"E, 62°34'59"N) tai (29.750000 E, 62.583332 N). Koska aistimaailmamme on kolmiulotteinen, myös kolmannen koordinaatin eli korkeuden käyttö on joissakin tapauksissa perusteltua. Kuvassa 3 hävittäjäkoneen paikka on luonnollisinta esittää kolmella koordinaatilla. *Ominaisuustietoa* voivat olla muun muassa tunnisteet, mittaushavainnot, luokitukset ja kuvailut. Esimerkiksi aiemmin mainittuun sanomalehteen voisi liittyä tieto lehden nimestä, ilmestymispaikasta tai sivumäärästä.



Kuva 3. Lentokoneen paikantamisessa tarvitaan kolmea koordinaattia.

Yksittäisellä koordinaattiparilla voidaan ilmaista vain yksittäistä pistettä maan pinnalla. Usein on kuitenkin tarpeen kuvata myös viivamaisia kohteita. Tällainen määrittely tehdään käyttäen useita peräkkäisiä koordinaattipareja, jotka kuvaavat *murtoviivan* taitepisteitä. Käyttäjän paikantamiseen perustuvissa sovelluksissa voidaan tallentaa kulkupisteet, joiden kautta henkilö on kulkenut. Kun nämä pisteet yhdistetään saadaan reitti, joka on murtoviiva. Reaalimaailmassa viivamaisina kohteina voivat olla esimerkiksi tiet, joet tai puhelinlinjat.

Pistemäisten ja viivamaisten kohteiden lisäksi paikkatieto voi esiintyä aluemaisina kohteina. Epäsäännöllisen muotoisten kohteiden ulkorajana toimii murtoviiva. Kuvassa 4 Suomen kartalta on murtoviivalla rajattu Itä-Suomen alue. Säännöllisen muotoilla alueilla rajana voivat toimia geometriset muodot, kuten ympyrä tai neliö. Aluemaisina kohteina voidaan käsitellä esimerkiksi kaupunkeja, rakennuksia tai hallintoalueita.



Kuva 4. Itä-Suomen alue rajattuna Suomen kartalta.

Internetiä voidaan pitää maailman suurimpana lähes tutkimattomana paikkatiedon lähteenä [12]. Internetin voimakas leviäminen sai alkunsa vuonna 1989, kun Tim Berners-Lee keksi CERN:in laboratoriossaan internetiin perustuvan hypermedian jakelukanavan, *World Wide Web:in* (*WWW*) [10]. Tietoverkon selaaminen graafisen selaimen avulla tekee kynnyksen verkon käyttämiseen matalaksi. WWW koostuu pääasiassa *HTML-kielellä* (*Hypertext Markup Language*) kuvatuista hypertekstidokumenteista eli WWW-sivuista. Tyypillinen sivu sisältää tekstiä, kuvia sekä hyperlinkkejä, jotka viittaavat toisiin WWW-dokumentteihin. Lisäksi WWW-sivuilla voi olla muun muassa Java-sovelluksia, videoita, pdf-dokumentteja tai melkein mitä tahansa sähköisessä muodossa olevaa multimediaa. Perinteisemmät WWW-palvelut säilyttävät vanhaakin tietoa arkistomateriaalina saatavilla ja uusia palveluita ja informaatiota syntyy koko ajan lisää. Esimerkiksi verkon suosituin hakukone Google suorittaa hakunsa vuonna 2004 yli 8 miljardin sivun joukosta [5].

Normaalisti tiedon etsiminen verkosta aloitetaan valitsemalla sopivat avainsanat, jotka syötetään hakukoneeseen. Hakukone etsii sivuja, joista halutut sanat löytyvät. Tämän jälkeen hakukone asettaa löytyneet sivut jollakin tavalla paremmuusjärjestykseen. Tästä hakutuloksesta käyttäjä valitsee itselleen *relevanteimmat*, eli hakukriteeriä parhaiten vastaavat tulokset. Mikäli haku ei tuottanut tyydyttäviä tuloksia, käyttäjä valitsee uudet hakusanat ja suorittaa haun uudelleen.

Nykyaikaiset hakukoneet toimivat erittäin kattavasti ja melko luotettavasti. Herää kuitenkin kysymys, voisiko hakutulosten rankkausta vielä parantaa. Otetaan esimerkiksi turisti, joka WAP-puhelimellaan etsii verkosta tietoa pizzerioista. Mikäli haku suoritetaan Joensuun keskustassa, olisi järkevää, että hakukone osaisi automaattisesti rankata joensuulaisten pizzerioiden verkkosivut korkeammalle kuin kaukaisemmat vastineensa.

Mobiileissa päätelaitteissa nykyisten verkkopalveluiden käyttäminen on hankalaa johtuen rajallisista tiedonsyöttö- ja näyttömahdollisuuksista. Muun muassa matkapuhelimen näppäimistö ei ole kovinkaan hyvä väline tekstien syöttämiseen. Jos hakukoneeseen pystyisi syöttämään lyhyesti sen mitä haluaa hakea ja hakukone ottaisi lisäksi huomioon käyttäjän sijainnin määrittäessään relevanteimpia hakutuloksia, olisi verkon käyttäminen mobiileilla päätelaitteilla paljon hyödyllisempää kuin nykyisin.

Paikkatiedon esiintymistiheys on merkittävä tekijä päätettäessä, kannattaako uusia sovelluksia rakentaa netistä löytyvän paikkatiedon varaan. Mitä enemmän paikkatietoa voidaan löytää sitä paremmin sitä voidaan hyödyntää. Tässä tutkielmassa olen tutkinut, kuinka paljon suomalaiset verkkopalvelut sisältävät paikkatietoelementtejä. Luvussa 2 tutkiskelen tavallisimpia tilanteita, joissa paikkatietoa esiintyy HTML-dokumenteissa sekä teksteissä yleisesti ja esittelen, kuinka paikkatieto voidaan luokitella sen luotettavuuden mukaisesti eksaktiin, epäsuoraan ja pääteltävään paikkatietoon. Koska on käynyt ilmi, että eksaktia koordinaattimuodossa olevaa ja helposti kerättävää paikkatietoa on netissä varsin vähän, olen keskittynyt epäsuoran paikkatiedon keräämiseen. Epäsuoran paikkatiedon kohdistamiseen läheisesti liittyviin paikkatietorekistereihin luon lyhyen katsauksen luvussa 3. Luvussa 4 kerron kuinka paikkatiedon eristäminen nettidokumenteista voidaan käytännössä toteuttaa. Lisäksi esittelen keräämästäni aineistosta muodostamani tutkimustulokset. Luvussa 5 teen yhteenvedon tutkielmasta sekä esitän päätelmiä aiheeseen liittyen.

2 PAIKKATIEDON ESITYSTAVAT TEKSTIDOKUMENTEISSA

Monet www-sivut ovat yritysten ja yhteisöjen ylläpitämiä. Niissä annetaan tietoa yritysten palveluista ja tuotteista. Kuinka tällaisen dokumentin sitten pystyisi liittämään sijaintitiedon? Sivusto sisältää hyvin harvoin tarkkaa koordinaattitietoa. Usein sijaintitieto on kerrottu osoitetietona tyyliin *Koulukatu 117, 80100 Joensuu*. Tästä tiedetään, että yritys sijaitsee Koulukadulla Joensuussa ja paikantaminen onnistuu varsin helposti, jos käytössä on osoiterekisteri.

Toinen yleinen yhteystieto on puhelinnumero. Tavallisen lankapuhelinnumeron karkea paikantaminen onnistuu suuntanumeron perusteella. Tarkkaan paikantamiseen tarvitaan sähköistä puhelinluetteloa, joka annetulla numerolla haettaessa palauttaa liittymän omistajan nimen ja osoitteen. Sama keino sopii suuntaa-antavasti myös matkapuhelinliittymän paikantamiseen.

Suuremman haasteen tarjoavat tekstit, joissa yhteystietoja ei suoraan kerrota. Esimerkiksi lehtiartikkelit, jotka kohdistuvat usein johonkin paikkaan ja tämä paikka kerrotaan tekstissä. Esimerkiksi *auto törmäsi puuhun Väärälammentiellä* tai *presidentti avasi Kuusamon kesäjuhlat*. Ongelman aiheuttavat suomen kielen taivutusmuodot, joita on varsin paljon ja paikannimet voivat esiintyä missä tahansa muodossa. Suomea äidinkielenään puhuva henkilö tunnistaa helposti, että sana *Huittisissakin* viittaa *Huittinen*-nimiseen paikkakuntaan, mutta taivutusmuotojen tunnistaminen automaattisesti on vaikeampaa.

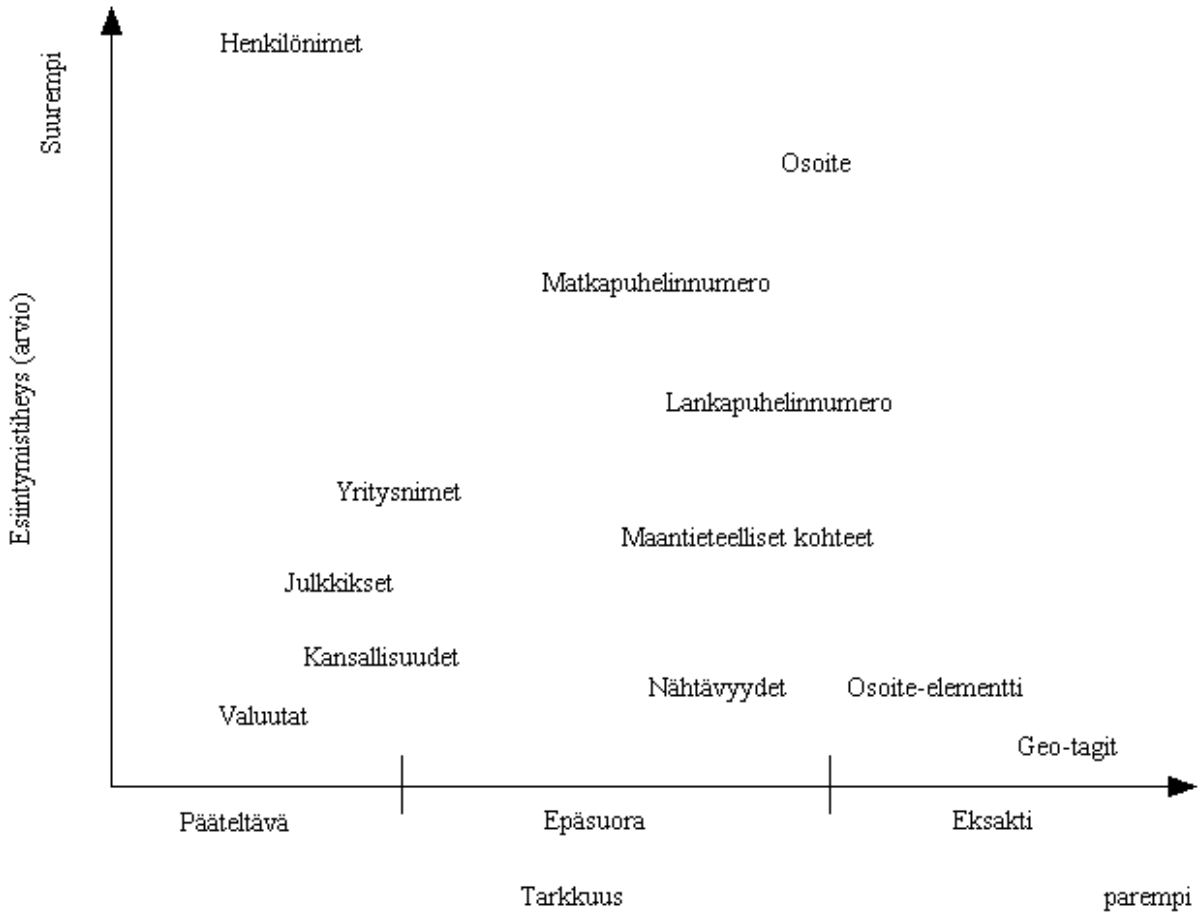
Varsinaisten paikannimien lisäksi lehtiartikkelissa voi olla jotakin muuta tietoa, jolla dokumentin voi paikantaa. Esimerkiksi eduskuntatalo on paikka, jonka jokainen suomalainen tietää sijaitsevan keskellä Helsinkiä. Muita vastaavanlaisia sanoja ovat tietyn paikkakunnan asukasta tai tuotetta kuvaavat adjektiivit taivutusmuotoineen kuten *rovaniemeläinen* tai *suomalainen*. Ulkomaihin kohdistuvia tekstejä voi lisäksi paikantaa karkeasti, jos tekstissä vilahtaa yksikäsitteinen valuutan nimi. Esimerkiksi sana *rupla* viittaa vahvasti Venäjään.

Osoitteet, puhelinnumerot, nähtävyydet ja luonnonkohteet sisältävät siis paikkatietoa, mutta eivät suoraan kerro kohteen eksaktia sijaintia. Käytän tällaisesta paikkatiedosta nimitystä *epäsuora paikkatieto*. *Pääteltävää paikkatietoa* sisältävät ilmaukset, jotka kohdistuvat yleensä suurelle alueelle tai moniselitteisesti useille alueille. Tällaisissa tapauksia ovat muun muassa valuutat tai kansallisuudet, joiden maantieteellinen sijoittuminen on pääteltävissä. Pääteltävä paikkatieto on luotettavuudeltaan epätarkempaa kuin epäsuora paikkatieto. Kohdistettaessa epäsuoraa paikkatietoa

tarvitaan avuksi *paikkatietorekisteri*, joka sisältää tiedon kohteesta sekä sen koordinaatit. Hyvänä esimerkkinä arkipäiväisestä paikkatietorekisteristä voisi olla jo aiemmin mainittu puhelinluettelo, jonka avulla puhelinnumeron voi paikantaa johonkin osoitteeseen.

Muita menetelmiä web-dokumentin paikantamiseen ovat esimerkiksi palvelimien IP-osoitteisiin perustuva paikantaminen ja hyperlinkkirakenteesta arvioitu paikantaminen. IP-osoitteen perusteella voidaan karkeasti paikantaa sen palvelimen fyysinen sijainti, jossa web-sivusto sijaitsee. Hyperlinkkipohjainen paikantaminen perustuu siihen, että sivut, jotka ovat yhteydessä toisiinsa linkillä, voivat paikantua samaan paikkaan. IP-osoitteisiin tai hyperlinkkeihin perustuvat menetelmät voivat olla kuitenkin varsin epätarkkoja tapoja paikantaa web-dokumentteja.

Kuvassa 5 on vertailtu erilaisia sijaintitietoa sisältäviä elementtejä. Mitä lähempänä kuvaajan oikeaa reunaa elementti on, sitä tarkempaa on sen sisältämä sijaintitieto. Mitä korkeammalla kuvaajasta elementti löytyy, sitä useammin verkkosivustoilta löytyy kyseisiä ilmauksia. Kuvaaja perustuu omaan arviooni, mutta jonkinlaista käsitystä eri elementtien sijoittumisesta toisiinsa nähden se tarjoaa. Mitä lähempänä oikeaa yläkulmaa elementti on, sitä käyttökelpoisempi se on. Kuten huomataan, geo-tagit tarjoavat tarkimman sijainti-informaation, mutta niitä käytetään hyvin vähän. Toisaalta henkilönimiä löytyy eniten, mutta niiden yksiselitteinen paikantaminen on vaikeaa. Epäsuorasta sijaintitiedosta johdetun eksaktin sijaintitiedon tarkkuus riippuu siis suuresti käytettävästä tiedon lähteestä [13]. Kuvaajasta käy myös ilmi jaottelu pääteltävän, epäsuoran ja eksaktin sijaintitiedon välillä.



Kuva 5. Sijaintitietoa sisältävien elementtien tarkkuus ja esiintymistiheys tekstidokumenteissa.

Tekstidokumentteja voidaan tuottaa sähköiseen muotoon useilla eri tavoilla. Tavallisimmillaan tekstin tuottamiseen käytetään nykyisin kehittyneitä tekstinkäsittelyohjelmia. Eräs modernien tekstinkäsittelyohjelmien perusedellytyksistä on niin sanottu *wysiwyg*-periaate (*What you see is what you get*). Tämä tarkoittaa sitä, että ohjelma hoitaa tekstin muotoilutietojen tallentamisen ilman, että käyttäjä sitä varsinaisesti huomaakaan. Näin ollen kirjoittajaa kiinnostaa vain tekstin sisältö, ei juurikaan se, että hän jotenkin erikseen ilmaisisi dokumentin sijaintitiedon.

WWW-maailmaan tekstiä tuotetaan yleensä HTML:llä [10]. Se on tagi-pohjainen merkkauskieli, jolla kirjoittaja itse määrittelee, miltä teksti näyttää. *Tagi* on merkintä, jolla dokumentin kirjoittaja tietoisesti kertoo dokumenttia käsittelevälle sovellukselle millaista tietoa teksti sisältää. dokumenttia käsittelevä sovellus (yleensä nettiselain) tietää näiden merkintöjen perusteella muotoilla tekstin ulkoasun halutuksi. HTML-tagien avulla kirjoittaja voi vaikuttaa lisäksi muun muassa sivun asetteluun ja linkkirakenteeseen.

On olemassa myös sellaisia HTML-editoreita, jotka toimivat wysiwyg-periaatteella. Tekstin muotoilemisen ohella kirjoittaja on usein myös kiinnostunut tekstin ulkopuolisista asioista, kuten dokumentin näkymisestä hakukoneille. Tällaisiin tarkoituksiin käytetään HTML:n META-elementtiä, jolla voidaan määritellä esimerkiksi dokumentin kirjoittaja, avainsanat ja muita vastaavia tietoja. META-elementtiä voidaan myös käyttää dokumentin sijaintitiedon määrittämiseen [13].

Kuten edellä tuli mainittua, tekstin kirjoittaja harvemmin merkitsee dokumenttiin sijaintitietoa siten, että se olisi helposti käytettävissä. Tämä aiheuttaa sen, että epäsuoraa paikkatietoa sisältävät ilmaukset tulisi pystyä tunnistamaan ja keräämään tekstirungosta automaattisesti.

2.1 Eksakti sijaintitieto

HTML on *World Wide Web Consortiumin* [15] (*W3C*) standardoima merkkauskieli, jota käytetään hypertekstin julkaisemiseen WWW:ssä. Uusin HTML:n versio 4.01 on standardoitu vuonna 1999. Uudempi menetelmä olisi *Extensible Hypertext Markup Language* (*XHTML*), joka on laajennus HTML:stä. HTML on silti vielä nykyisin valtakieli luotaessa WWW-sivuja [10].

Standardi HTML ei tarjoa juurikaan mahdollisuuksia sijaintitiedon esittämiseen muuten kuin tekstisisällössä. Ainoa standardin mukainen tapa on käyttää ADDRESS-tagia, joka kertoo, että kysymyksessä on osoitetieto. Standardin ulkopuolisia tapoja on useampiakin, joista myöhemmin esittelen Geo-tagi -menetelmän. Standardin ulkopuoliset tavat eivät ole saavuttaneet laajempaa käyttäjäkuntaa.

2.1.1 Geo-tagit

Geo-tagit (*geotags*) ovat syntyneet *The Internet Engineering Task Force*:n (*IETF*) *Geographic registration of HTML documents*-projektin tuloksena [3][27]. IETF on avoin organisaatio, joka koostuu muun muassa verkkosuunnittelijoista, ohjelmoijista ja tutkijoista, joiden yhteisenä tavoitteena on parantaa internetin toimivuutta ja arkkitehtuuria. Geo-tagi -projekti on tätä kirjoitettaessa kesken ja muutoksia tageihin saattaa siten seuraavassa esitettyyn nähden vielä tulla.

Geo-tagi -projektin lähtökohtana on ollut helpottaa verkkosivujen paikantamista. HTML-sivut viittaavat usein johonkin paikkaan maan pinnalla. Geo-tagit mahdollistavat maantieteelliset haut, joita varten nykyisiä hakusanoihin perustuvia hakukoneita ei ole suunniteltu [3]. Geo-tagi-

menetelmä mahdollistaa staattisen sijaintitiedon lisäämiseen tavallisiin HTML-dokumentteihin käyttäen META-elementtiä. WWW-sivujen tekijöillehän META-elementti on yleensä entuudestaan tuttu, eikä sen käyttö aiheuta ongelmia.

Koska useimmilla WWW-sivujen tekijöille paikkatietojärjestelmien terminologia ei kuitenkaan ole jokapäiväisessä käytössä, on geo-tageista yritetty tehdä mahdollisimman helppokäyttöisiä. Ne sisältävät mahdollisimman pienen joukon käytettäviä tunnisteita [3]. Näitä tunnisteita on kolme: *geo.position*, *geo.region* ja *geo.placename*. Geo.position-tunnistetta käytetään pituus- ja leveysasteiden sekä korkeuden määrittelyyn. Geo.region on tunniste, jota käytetään alueen määrittämiseen. Alue tulee määritellä ISO 3166-2-standardin mukaan [7]. Vapaaehtoista geo.placename-tunnistetta käytetään, jos vielä paikannimikin halutaan ilmaista.

Geo.position-tunnisteella merkitään dokumentin sijaintitieto eksaktisti koordinaattien avulla. Geo-tagien koordinaattijärjestelmä koostuu kolmesta koordinaatista. Pituus- ja leveysasteet pyritään ilmoittamaan alle kilometrin tarkkuudella ja korkeus 25 metrin tarkkuudella [3]. Koska korkeusmittauksia suoritetaan eri tavoin, ei korkeustiedon käyttöä vielä suositella.

Geo.region-tunnisteessa käytettävä ISO 3166-2 standardi määrittelee kaikkien maiden eri alueille yksikäsitteiset tunnisteet [7]. Esimerkiksi Kanadan Quebecin tunniste on CA-QC. Jos tarkkaa aluetta ei tiedetä tulee käyttää pelkkää kaksikirjaimista maatunnistetta [3]. Mikäli HTML–dokumentti ei paikannu yksikäsitteisesti yhteen pisteeseen maan päällä, tulee geo.region tunnistetta käyttää yksinään.

Paikannimitunnistetta geo.placename tulee käyttää paikannimen selventämiseen. Tunnisteeseen voi lisätä vielä tiedon siitä, millä kielellä paikannimi on kirjoitettu [3]. Esimerkiksi paikannimeen Lontoo kannattaa liittää tieto siitä, että nimi on kirjoitettu suomeksi.

Kuvan 6 esimerkissä A kuvataan paikkaa, joka sijaitsee 4817 metriä meren pinnan yläpuolella pisteessä, joka on 45,92 astetta pohjoista leveyttä ja 6,92 astetta itäistä pituutta. Kyseinen paikka on Euroopan korkeimman vuoren Mont Blancin huippu. Metadataprofiilia ”http://geotags.com/geo” käytetään määrittelemään, että käytössä ovat juuri geo-tagit.

Esimerkissä B sijaintitieto on kuvattu pituus- ja leveysasteina ilman korkeuskoordinaattia. Esimerkissä dokumentti kohdistuu paikkaan, jonka sijainti on 49 astetta eteläistä leveyttä ja 7 astetta itäistä pituutta. Kuten huomataan, koordinaatit voidaan ilmaista myös ilman desimaaliosaa.

Esimerkeissä C ja D esitellään, kuinka geo.region-tunnisteen avulla paikka voidaan yksikäsitteisesti määritellä, jos samannimisiä paikkoja on olemassa useampia. Esimerkissä C kyseessä on Albert-niminen paikka Ranskassa, kun taas esimerkissä D käsitellään samannimistä paikkaa Saksassa.

HTML-attribuuttia ”lang” käytetään geo.placename-tunnisteen käytön yhteydessä, kun määritellään millä kielellä paikannimi on kirjoitettu. Esimerkissä E Lontoo on kirjoitettu ranskaksi.

Esim		Kartta
A	<head profile="http://geotags.com/geo”> <meta name="geo.position" content="45.92;6.92;4807">	
B	<meta name="geo.position" content="49;7">	
C	<meta name="geo.placename" content="Albert"> <meta name="geo.region" content="FR">	
D	<meta name="geo.placename" content="Albert"> <meta name="geo.region" content="GM">	
E	<meta name="geo.placename" lang="fr" content="Londres">	

Kuva 6. Esimerkkejä geo-tagien käytöstä.

2.1.2 Osoite–elementti

HTML:ssä perusmenetelmä osoitteiden merkkaamiseen on address-elementti. Se koostuu alku- (<ADDRESS>) ja lopputagista (</ADDRESS>), joiden väliin osoitetieto tulee [10]. Kun osoitetieto on tällä tavalla merkitty dokumenttiin, on sen tunnistaminen vaivatonta. Address-elementin ongelma on, että se sisältää myös muotoilutietoa. Oletuksena address kursivoi tekstin, jonka se sisältää. Niinpä sitä käytetäänkin lähinnä muotoiltaessa osoitetietoa ja toisaalta monet käyttäjät vierastavat sitä nimenomaan siksi, että se muotoilee tekstiä. Tämän tutkielman kokeellinen osuus osoittaakin, että address-elementti on varsin vähän käytetty menetelmä osoitteita muotoiltaessa, yleisempää on muokata osoitteen tyyli jotenkin muuten.

Address-elementillä merkittyä osoitetta voi pitää eksaktina johtuen kirjoittajan tietoisesta valinnasta erottaa osoite omaksi lohkokseen dokumentissa. Address-tagein merkityn osoitteen poimiminen HTML-muotoillusta tekstistä on huomattavasti helpompaa ja vähemmän virhealtista kuin merkkaamattoman osoitteen tunnistaminen. Tämä johtuu siitä, että pystytään täsmälleen tietämään mitkä tekstin osat kirjoittaja on halunnut sisältyvän osoitteeseen.

2.2 Epäsuora sijaintitieto

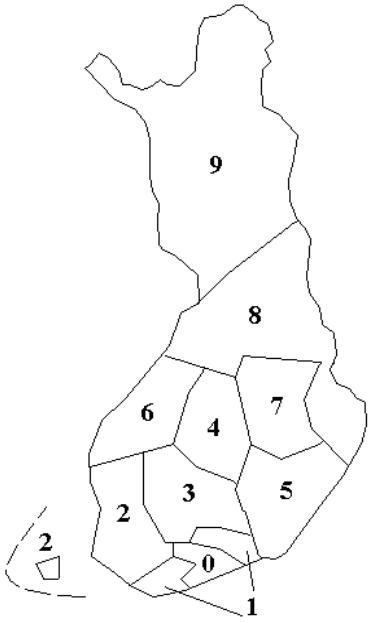
Tekstin sisältämä sijaintitieto esiintyy yleensä epäsuorana sijaintitietona. Web sisältää paljon tietoa, jolla on jonkinlainen maantieteellinen luonne. Erityisesti osoitteet, postinumerot ja puhelinnumerot ovat tällaista tietoa. Kaupallisessa tarkoituksessa ylläpidetyt verkkopalvelut sisältävät tyypillisesti yhteystiedot osoitteineen ja puhelinnumeroineen siinä missä tietoa tarjotuista tuotteista ja palveluistakin [16]. Verkossa ilmestyvien sanomalehtien artikkeleissa ilmaistaan yleensä myös paikka, jossa raportin kuvaamat tapahtumat ovat tapahtuneet. On siis luonnollista olettaa, että teksteillä on jonkinlainen yhteys tapahtuman paikkaan.

Tarkkaa tietoa siitä kuinka suuri osa web–sivuista sisältää epäsuoraa sijaintitietoa ei ole saatavilla. Varsin suureen web–materiaaliin perustuvan yhdysvaltalaistutkimuksen mukaan noin 4,5% kaikista sivuista sisältää tunnistettavan yhdysvaltalaisen postinumeron, 8,5% sisältää tunnistettavan puhelinnumeron ja 9,5% sisältää ainakin toisen näistä [13]. Tämä tulos on oikeastaan vain jonkinlainen alaraja sijaintitiedon määrälle. Tutkijat ovat päässeet tähän tulokseen omalla parserillaan, joka ei sisällä kaikkia mahdollisia keinoja epäsuoran sijaintitiedon tunnistamiseen.

2.2.1 Postiosoitteet ja postinumerot

Ehkä ilmeisin epäsuoran sijaintitiedon lähde ovat postiosoitteet. Ne ovat vuosisatojen aikana kehittyneet mahdollistamaan perinteisen kirjepostin jakelun tarkasti haluttuun paikkaan kaikkialla maailmassa. Postiosoitteiden automaattinen tunnistaminen tekstistä on varsin hyvin hallinnassa oleva ongelma, mutta se, että postiosoitteiden muotoilu vaihtelee suuresti eri maissa tekee asiasta monimutkaisen [13]. Lisäksi tietyn maan sisälläkin voi olla eroja osoitteen ilmaisemisessa. Kuitenkin useissa maissa postiosoitteiden ja postinumeroiden tunnistaminen luonnollisesta kielestä on määritelty tarkasti. Esimerkiksi Yhdysvalloissa postiosoitteiden paikantaminen onnistuu *Tiger/Zip+4*–osoiterekisterin avulla [16]. Suomessa postinumerot on määritelty yksiselitteiksi viisinumeroisiksi merkkijonoiksi [22].

Suomen postinumerot on jaettu maantieteellisesti alueittain. Kuvassa 7 on esitelty karkea aluejako, kuinka jako on suoritettu postinumeron ensimmäisen numeron perusteella. Esimerkiksi numerolla 4 alkavat postinumerot löytyvät Keski-Suomesta Jyväskylän seudulta. Ahvenanmaa kuuluu samaan postinumeroalueeseen Turun ja Porin kanssa.



Kuva 7. Suomen postinumeroalueet [22].

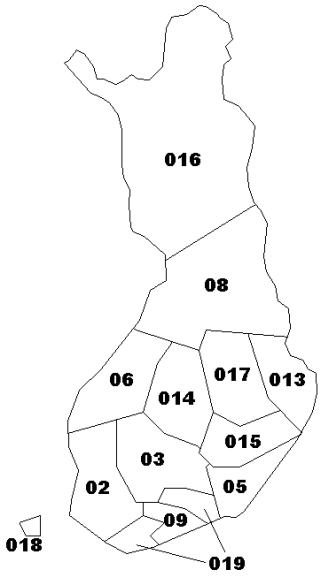
Normaalisti postiosoitteet on jaettu useisiin kenttiin, jotka erotetaan toisistaan pilkulla tai rivinvaihdolla. Tällaisia kenttiä ovat yleensä katuosoite, katunumero, postinumero, postitoimipaikka ja maa. Postiosoitteet voivat olla suhteellisia, jolloin lähettäjä tietää että vastaanottaja on samassa maassa, jolloin maan nimeä ei tarvita [13]. Inhimillisistä syistä johtuen osoitteet sisältävät myös kirjoitusvirheitä ja muotoilueroja. Satunnaisen osoitetiedon parsimisen suorittavan ohjelmiston toteuttaminen on yllättävän monimutkaista. Inhimillisten tekijöiden lisäksi ongelmia aiheuttavat monenlaiset lyhenteet, välimerkit ja muut vastaavat variaatiot.

2.2.2 Puhelinnumerot

Lankapuhelinnumerot on perinteisesti ryhmitelty suurimmaksi osaksi maantieteellisin perustein [16]. Syynä tähän on se, että tehokas puhelinliikenteen reitittäminen onnistuu parhaiten tällä tavoin. Tietenkin myös tällä säännöllä on lukuisia poikkeuksia. Kansainvälisessä puhelinliikenteessä puhelinnumerot alkavat tyypillisesti maakoodilla, joka sinällään mahdollista hyvin karkean maantieteellisen paikantamisen [13]. Mitä enemmän alueellisten puhelinnumeroiden rakenne voi tarjota tietoa, sitä tarkemmin paikantaminen voi onnistua.

Teksteissä esiintyy usein myös muita numeroita, jotka voi helposti tulkita puhelinnumeroiksi, vaikka ne ovatkin jotain muuta. Tällaisia ovat muun muassa sarjanumerot ja muut vastaavat pelkistä numeroista koostuvat merkkijonot. Tästä johtuu, että puhelinnumeroiden tunnistaminen ei voi onnistua täysin virheettömästi [13]. Kaikesta huolimatta on olemassa useampiakin standardeja tapoja kirjoittaa puhelinnumero, riippuen siitä onko tarkoituksena tarjota numeroa kansainväliseen vai kotimaan liikenteeseen. Useimmissa tapauksissa kansainväliset puhelinnumerot on *muotoiltu ITU-T*:n [8] (*International Telecommunication Union – Telecommunication Standardization Sector*, tunnettu aiemmin nimellä *CCITT*, *Comité Consultatif International Téléphonique et Télégraphique*) määrittelemän standardin mukaisesti. Numero alkaa + -merkillä, jota seuraa maakoodi, jonka jälkeen tulee aluekoodi. Joissain maissa aluekoodi voi olla myös kaupunkikohtainen. Maa- ja aluekoodit voivat olla erimittaisia, mutta koska niitä käytetään reititystarkoitukseen, ne ovat etumerkittömiä ja yksiselitteisiä. Siis maan ja alueen tunnistaminen puhelinnumerosta on varsin ongelmatonta.

Suurin osa puhelinliikenteessä koostuu kotimaanpuheluista. Tämän takia onkin tavallisempaa, että puhelinnumerot ilmaistaan mieluummin silmälläpitäen kotimaisia tarpeita, kuin kansainvälisiä standardeja [13]. Suomi on jaettu kolmeentoista telealueeseen, joista jokaisella on oma aluekoodinsa, joka tunnetaan paremmin suuntanumeron nimellä [23]. Kansainvälisiin tarkoituksiin puhelinnumero ilmaistaan muodossa +358 13 123 456, jossa +358 on Suomen maakoodi ja 13 Pohjois-Karjalan aluekoodi. Kotimaan puheluissa suuntanumero ilmaistaan merkitsemällä aluekoodin eteen 0, jolloin numero tyypillisesti kirjoitetaan 013 – 123 456. Paikallisliikenteessä puhelinnumeroon ei tarvita edes suuntanumeroa vaan pelkkä runkonumero, 123 456. Tällaisen numeron täysin varma paikantaminen on usein mahdotonta, koska sama runkonumero voi esiintyä useilla teleliikennealueilla. Kuvassa 8 on esitelty Suomen lankapuhelinverkon teleliikennealueet.



Kuva 8. Suomen teleliikennealueet [23].

Yksittäisen puhelinnumeron tasolle mentäessä on olemassa tietokantoja, jotka tarjoavat fyysisen osoitteen suurimalle osalle puhelinnumeroista. Tällaiset tietokannat ovat yleensä jonkin yrityksen omistuksessa ja suhteellisen kalliita loppukäyttäjälle, johtuen niiden kaupallisesta arvosta ja ylläpitokustannuksista [13].

2.2.3 Maantieteelliset kohteet

On olemassa paljon maantieteellisiä kohteita, joille varsinaista osoitetta ei ole määritelty, mutta joiden yksikäsitteinen paikantaminen onnistuu varsin helposti. Tällaisia ovat muun muassa vaarat, tunturit, joet ja järvet, joita Suomestakin löytyy tuhansittain. Lisäksi on paljon nähtävyyksiksi luokiteltavia paikkoja, jotka ovat yleisesti tunnettuja. Kansainvälisesti ajatellen esimerkiksi Eiffel-torni viittaa Pariisiin ja Akropolis Ateenaan. Suomessa Olavinlinna on tunnetusti Savonlinnassa ja eduskuntatalo Helsingissä.

Yleisesti saatavilla on joitakin tietokantoja, joissa tällaisia maantieteellisiä kohteita on listattu. Esimerkiksi *National Geospatial Intelligence Agencyn* [18] (*NGA*) ylläpitämä *Geonet Names Server* (*GNS*) sisältää maailmanlaajuisesti noin 5,5 miljoonaa paikannimeä koordinaattitietoineen [17]. Suomesta tietokannassa on yli 50 000 paikannimeä. Tietokannassa on monia kategorioita, kuten järviä, kouluja, kirkkoja ja rakennuksia. Lisätietoa GNS:stä löytyy luvusta 3.1.

2.3 Pääteltävä sijaintitieto

Tekstistä voi löytyä sanoja ja ilmauksia, joilla on selvästikin jonkinlainen maantieteellinen luonne, mutta joiden tarjoama sijaintitieto yleensä kohdistuu suurelle alueelle tai moniselitteisesti useisiin paikkoihin. Tällaisia sanoja ovat muun muassa kansallisuuksia, valuuttoja ja yritysnimiä kuvaavat sanat, joista seuraavassa hieman tarkemmin.

2.3.1 Kansallisuudet ja valuutat

Kansallisuudet ja monet valuutannimet viittaavat johonkin maantieteelliseen alueeseen. Kansallisuutta tai paikkakunnan asukasta tarkoittavien sanojen tunnistaminen tekstistä on hieman monimutkaisempaa kuin normaalien paikannimien, koska suomen kielessä kansallisuudet tulee kirjoittaa pienellä alkukirjaimella. Toisaalta tämä käytäntö ehkäisee virheitä, jotka aiheutuvat siitä, että sukunimi tulkitaan kansallisuudeksi. Esimerkiksi sukunimi Ruotsalainen ei viittaa Ruotsiin vaan varsin vahvasti Suomeen. Kansallisuudet tai paikannimistä johdetut asukasta tarkoittavat adjektiivit voidaan yleensä paikantaa varsin karkeasti ehkä vain kokonaisen valtion alueelle.

Valuuttojen nimiä hyväksi käyttäen teksti voidaan joissakin tapauksissa kohdentaa yksittäisen valtion tarkkuudella. Tämän menetelmän ongelma on, että samannimisiä valuuttoja on käytössä useissa maissa. Esimerkiksi euro on käytössä monissa EU-maissa ja Yhdysvalloilla, Kanadalla, Australialla ja Singaporella on omat dollarinsa. Vastaavasti sloty on käytössä ainoastaan Puolassa ja jeni Japanissa.

2.3.2 Muita menetelmiä

Yksi tavallisimmista karkean sijaintitiedon lähteistä on dokumentissa käytetty kieli. Ruotsinkielinen teksti sinällään ei tarkoita sitä, että teksti sisältäisi tietoa Ruotsista, mutta kielen ja maantieteellisen sijainnin välinen yhteys on silti melko korkea. Käytetyn kielen tunnistamisen voi tehdä monin eri tavoin, kuten HTML:n META–elementtien avulla tai suoralla kielianalyysillä.

Edellisissä aliluvuissa kuvattujen menetelmien lisäksi on olemassa vielä muitakin mahdollisia tapoja paikantaa tekstiä. Esimerkiksi puhelinluetteloa voitaisiin käyttää yksittäisten henkilönimien yhdistämiseen puhelinnumeroon ja tätä kautta osoitteeseen. Ongelma on luonnollisesti se, että henkilönimet eivät ole yksikäsitteisiä vaan miltei aina jokaiselle löytyy jostain kaima. Osaratkaisu voisi olla rekisteri, jossa määritellään rajattu määrä henkilönimiä, joille tarjotaan osoitetieto vaikkapa paikkakunnan tarkkuudella. Tällaiseen rekisteriin laskettaisi esimerkiksi kansanedustajat

ja nimekkäimmät taiteilijat sekä huippu-urheilijat. Perusteluna sille, että tällainen menetelmä voisi toimia on, että nimi *Paavo Lipponen* viittaa hyvin paljon todennäköisemmin nykyiseen eduskunnan puhemieheen, kuin johonkin hänen kaimaansa.

Myös yritysnimien paikantamista voisi yrittää esimerkiksi keltaisten sivujen avulla. Tämän menetelmän ongelmaksi muodostuvat useat haarakonttorit ja vastaavat, joiden takia yksikäsitteistä osoitetta yrityksille on monissa tapauksissa hyvin hankala määritellä.

2.4 Paikannimiin liittyviä semanttisia ongelmia

Kun nimi *Eiffel* esiintyy tekstissä, oletetaan luonnollisesti että tekstillä on jokin yhteys maailmankuuluun pariisilaiseen torniin. Tällä perusteella dokumentti voitaisi jo kohdistaa Pariisiin. Kuitenkin tässä tapauksessa on olemassa merkittävä virheen mahdollisuus, koska sanalla *Eiffel* on monia merkityksiä, joista torni on tunnetuin. Nimellä *Eiffel* tunnetaan muun muassa ohjelmointikieli, italialainen tanssiyhtye ja australialainen lääkeyhtiö. Näin ollen kun sana *Eiffel* tunnistetaan tekstistä, täytyy tekstin kohdistaminen varmistaa myös jollakin muulla tavalla [13].

Suomessa, kuten useissa muissakin kielissä, monet henkilönimet ovat kehittyneet maantieteellisten paikannimien mukaan. On myös mahdollista, että paikka on saanut nimensä lähiseudulla vaikuttaneen henkilön mukaan. Näistä syistä johtuen monille nimille löytyy maantieteellinen vastine, jolla ei ole mitään tekemistä itse henkilön kanssa. Esimerkiksi dekkarikirjailija Matti Yrjänä Joensuusta ei juurikaan voi vetää yhteyttä Joensuun kaupunkiin. Tästä ongelmasta seuraa se, että tavalliset henkilönimet voivat aiheuttaa tekstin kohdistamisessa satunnaisia virheitä, jotka voivat paikantua periaatteessa mihin päin Suomea tahansa.

Monet yritykset ovat saaneet nimensä paikkakunnan nimen mukaan. Tunnetuimpia tällaisia suomalaisyrityksiä ovat muun muassa Nokia, Outokumpu ja Wärtsilä. On luonnollista, että kohdattaessa tekstissä tällainen yritysnimi, dokumenttia ei voida kohdentaa vastaavalle paikkakunnalle. Yritysnimet voivat aiheuttaa dokumenttien paikantamisessa samanlaisia satunnaisvirheitä kuin henkilönimetkin.

Henkilönimien ja yritysnimien sekaantuminen voidaan varsin tehokkaasti välttää käyttäen *nimettyjen kohteiden tunnistusta* (*Named Entity Recognition, NER*) [1][28][29]. Tällaisilla menetelmillä henkilönimet, yritysnimet ja paikannimet pystytään tunnistamaan tekstistä tiettyjen

sääntöjen perusteella myös ilman paikkatietorekisteriä. Tämän tutkielman laajuus ei anna mahdollisuutta pureutua nimettyjen kohteiden tunnistamiseen tämän tarkemmin.

Vaikka tekstistä pystyttäisikin tunnistamaan sana *Joensuu* ja oltaisi vielä varmoja, että kyseessä on juuri pohjoiskarjalainen Joensuu, voi silti olla niin, että dokumentti kohdistuu jonnekin muualle kuin Joensuuhun [2]. Esimerkiksi tekstissä voi esiintyä ilmaus *noin 120 kilometriä Joensuusta länteen*, joka viittaa johonkin paikkaan Kuopion ja Varkauden välimaastossa. Toinen vastaavanlainen tapaus voisi olla *Joensuu-Vaasa – linjan eteläpuolella*. Dokumentin, joka sisältää tällaisen lausahduksen, tulisi kattaa koko eteläinen Suomi kyseisen linjan eteläpuolelta.

Monissa tapauksissa useita paikkoja voidaan epäsuorasti sitoa joukoksi, jota tekstissä on helpompi käsitellä, kuin yksittäisiä paikannimiä [2]. Esimerkiksi ilmaus *Pohjois-Karjalan kaupungit* sisältää muun muassa Joensuun, Nurmeksen ja Lieksan. Jos edelliseen vielä lisätään negaation sisältävä *paitsi Outokumpu*, on dokumentin tarkoituksenmukainen kohdistaminen jo varsin vaikea tehtävä.

Lauserakenteista ja sanojen sisällöistä johtuvien ongelmien ratkaiseminen sovellustasolla vaatisi varsin suuria ponnisteluja. Dokumentille tulisi tehdä varsin tarkka kielellinen analyysi, joka sisältäisi lauserakenteiden purkamisen sekä eri sanojen roolien ja merkitysten tunnistamisen. Artikkelissa [2] Bilhaut *et al*. esittelevät periaatteet, kuinka tällainen analyysi on suoritettu ranskankielisille teksteille.

3 Paikkatietorekisterit ja niitä hyödyntäviä web-sovelluksia

Paikkatietoelementtien kohdistaminen tiettyyn maantieteelliseen paikkaan onnistuu paikkatietorekisterien avulla. Paikkatietorekisteri voi sisältää koordinaattimuodossa olevan sijaintitiedon lisäksi paljon muutakin informaatiota riippuen rekisterin käyttötarkoituksesta. Maantieteellisten kohteiden rekisteri (engl. *gazetteer*) voi sisältää informaatiota hyvinkin erilaisista kohteista, eikä ole epätavallista, että samassa rekisterissä on tietoja maanosista ja valtioista aina kouluihin ja puistoihin asti [11]. Osoiterekisteri taas sisältää postiosoitteita ja niiden koordinaatteja. Muun muassa metsäteollisuudella, liikenteellä ja teleoperaattoreilla on omat paikkatietorekisterinsä, jotka sisältävät juuri niiden liiketoiminnalle olennaista tietoa.

Paikkatietorekistereiden kuten muidenkin vastaavien suurten tietokantojen hyötykäyttö on vaikeaa ilman sovellusta, jonka avulla käyttäjä voi suorittaa hakuja ja muita kyselyitä tietokantaan. Sovelluksia on olemassa monenlaisia riippuen rekisterin käyttäjien tarpeista. Sovellus voi myös yhdistellä eri rekistereiden tietoja ja muodostaa näin johdettua tietoa, jota yksittäiset rekisterit sinällään eivät tarjoa. Esimerkkinä tällaisesta tietojen yhdistämisestä ovat erilaiset reittipalvelut, jotka yhdistävät osoiterekisterin ja tierekisterin tietoja ja näiden perusteella muodostavat halutun reitin lähtöosoitteesta pääteosoitteeseen.

Seuraavissa aliluvuissa esittelen kaksi paikkatietorekisteriä sekä kaksi sovellusta, joiden toiminta on rakennettu paikkatietorekistereiden varaan. *GeoNet Names Server (GNS)* ja *Alexandria Digital Library Gazetteer* (*ADL Gazetteer*) ovat maantieteellisiä kohteita sisältäviä rekistereitä. *Keltaiset sivut* ja *Google Local* edustavat sovelluksia, joiden toiminta perustuu paikkatietorekistereihin.

3.1 GeoNet Names Server

GeoNet Names Server (*GNS*) on maantieteellisiä kohteita sisältävä varsin kattava paikkatietorekisteri [17]. Siitä löytyi syyskuussa 2004 noin neljä miljoonaa kohdetta ja yhteensä noin 5,5 miljoonaa paikannimeä. Paikannimien ylimäärä verrattuna kohteiden lukumäärään johtuu samaa paikkaa tarkoittavista synonyymeista ja erikielisistä nimityksistä. GNS mahdollistaa haut kahteen erilliseen paikkatietorekisteriin siten, että käyttäjä tuntee käyttävänsä vain yhtä rekisteriä. Käytettävät rekisterit ovat Yhdysvaltain sisäisistä tiedoista koostuva *National Geospatial-Intelligence Agencyn* (*NGA*) tietokanta sekä muiden maiden kohteita sisältävä *U.S. Board on Geographic Names*:in (*US BGN*) tietokanta. NGA toimii palvelun ylläpitäjänä.

NGA toimii Yhdysvaltain puolustusministeriön alaisuudessa tarjoten lakisääteistä lisätietoa paikallisille päätöksentekijöille ja muille valtiollisille toimielimille. NGA toimittaa mahdollisimman ajantasaista ja tarkkaa maantieteellistä informaatiota Yhdysvaltain alueelta [18]. BGN toimii NGA:n tavoin palvellakseen Yhdysvaltain toimielimiä. Sen tavoite on standardoida maantieteellisten nimien kirjoitusasuja Yhdysvaltain hallinnon julkaisuissa [17].

GNS-paikkatietorekisteristä on saatavilla joko kaikki kohteet kattava versio tai alueittaiset versiot. Aluekohtaisia paketteja on 247, jokaiselle valtiolle omansa sekä lisäksi esimerkiksi tietyille saariryhmille ja merenpohjan alueille omansa. Suomen alueelta rekisterissä oli syyskuussa 2004 yhteensä 53179 kohdetta.

Nimen ja koordinaattitietojen lisäksi GNS sisältää kohteista useita muitakin tietoja. Jokaisesta kohteesta rekisteriin on tallennettu 21 eri attribuuttia. Koordinaattitiedot on tallennettu sekä desimaalimuodossa että kokonaislukuna. Kohteet on luokiteltu yhdeksään eri luokkaan, joita ovat muun muassa hallinnolliset alueet, asutut paikat, viljelyalueet ja tiet. Nämä yhdeksän luokkaa on jaettu vielä tarkempiin aliluokkiin, joita rekisteristä löytyy yhteensä 643 erilaista. Tällaisia aliluokkia ovat esimerkiksi lentokentät, hotellit, rannat ja mäet.

Suomen alueella GNS:n sisältö rajoittuu erilaisiin paikannimiin ja hallinnollisiin alueisiin. Usein paikannimistä löytyvät sekä suomen- että ruotsinkieliset vaihtoehdot. Joissakin tapauksissa rekisteri sisältää myös saamen- ja venäjänkielisiä nimiä. Asutuskeskusten lisäksi rekisteristä löytyy paljon muun muassa vesistöjen, saarien, jokien ja soiden koordinaatteja. Hotelleita tai muita ihmisen rakentamia rakennuksia rekisteristä ei juurikaan Suomen alueelta ole.

Ladattavien rekisteritiedostojen lisäksi GNS tarjoaa myös käyttöliittymän hakujen tekemiseen rekisteriin. Hakuja voi tehdä paikannimeen perustuvan haun lisäksi kohteiden luokituksien perusteella. Hakutuloksena näytetään kaikki hakua vastaavat rekisterin rivit tekstimuodossa, mitään karttaan perustuvaa tulostusta ei tarjota. Muutenkin käyttöliittymä on vaivalloinen ja hidas. Esimerkiksi yksinkertaisella hakusanalla Joensuu haku kestää noin puoli minuuttia. Käyttöliittymän toteuttajien tarkoituksena lieneekin ollut vain tarjota mahdollisuus tutustua rekisterin sisältöön, ei niinkään se, että rekisteristä tehtäisiin helppokäyttöinen.

Kaiken kaikkiaan GeoNet Names Server tarjoaa kattavan ja laajan maantieteellisten kohteiden rekisterin, josta tarvittaessa löytää haluamansa kohteen koordinaatit. Rekisteritiedostoa voi pienellä

muokkauksella käyttää erilaisissa sovelluksissa, kuten itsekin olen tehnyt tämän tutkielman prototyyppisovelluksessa.

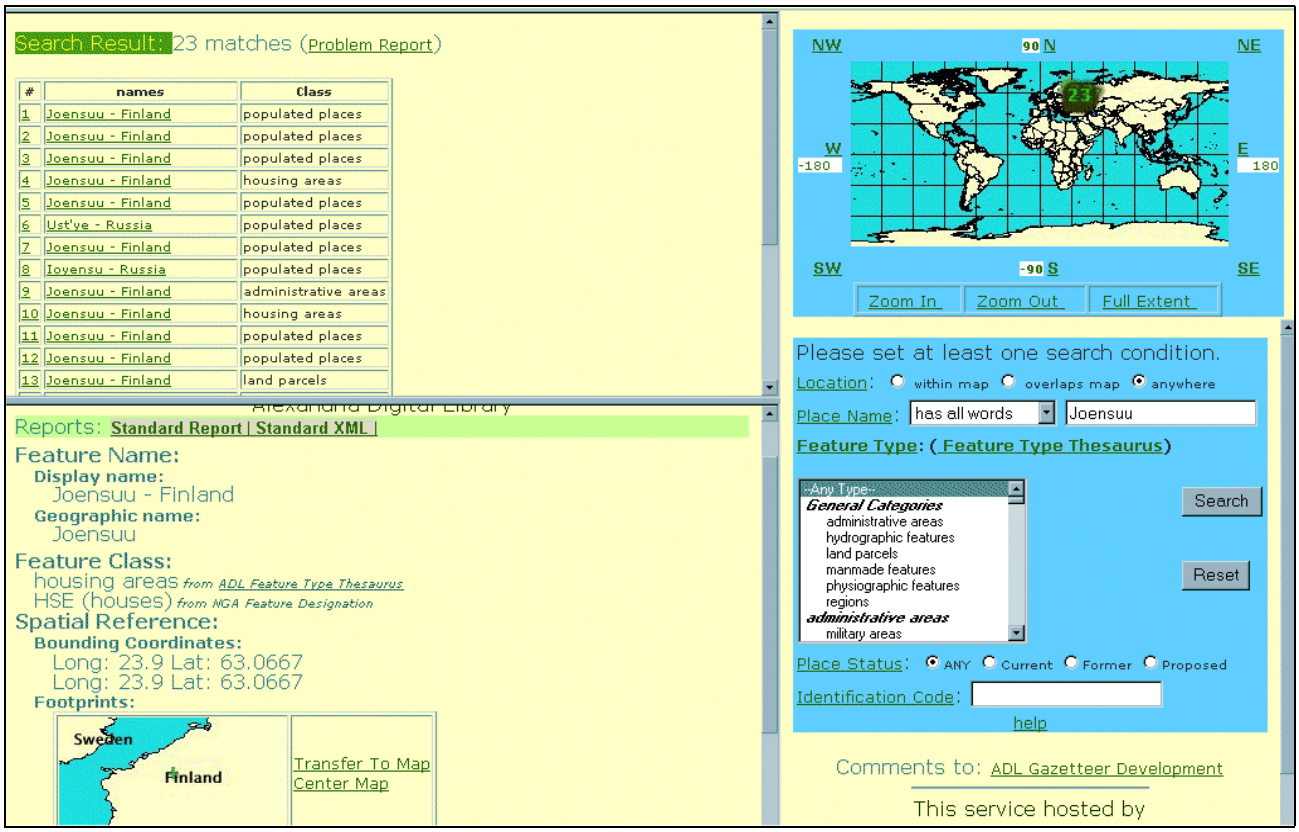
3.2 Alexandria Digital Library Gazetteer

Alexandria Digital Library (*ADL*) on kokoelma maantieteellisiin kohteisiin viittaavia materiaaleja, jotka on jaettu kaikkien verkon käyttäjien saataville [19]. ADL sisältää myös tutkijoiden käyttöön tehtyjä dokumentteja koskien digitaalisen kirjaston arkkitehtuuria, paikkatietorekisterisovelluksia, koulutukseen liittyviä sovelluksia ja muita ohjelmistokomponentteja. ADL tarjoaa myös HTML-käyttöliittymään kokoelmiinsa ja paikkatietorekisteriin.

ADL Gazetteer on paikkatietorekisteri, jonka tehtävänä on yhdistää maantieteelliseen nimiin maantieteellinen sijaintitieto ja muuta kuvaavaa informaatiota. Rekisteriä voidaan käyttää halutun paikan sijainnin etsimiseen tai tietyn alueen sisältämien nimettyjen paikkojen tutkimiseen. Hakuja voi tehdä myös rajattuihin kohdeluokkiin, kuten vesistöihin, tai jokiin. ADL Gazetteer yhdistelee tietonsa eri tietolähteistä, joista merkittävimmät ovat *U.S Geological Survey*:n *GNIS*-tietokanta ja edellisessä aliluvussa esitelty GNS-rekisteri. Näin ollen ADL:n paikkatietorekisteriä voidaan pitää laajempana kuin GNS-rekisteriä. ADL Gazetteer on perustettu vuonna 1999 ja sitä ylläpitää Kalifornian yliopisto Santa Barbarassa.

Kaikki ADL-rekisterin sisältämät paikannimet on koottu yhdeksi tiedostoksi, jonka käyttäjä voi ladata omaan käyttönsä. Tiedosto sisältää syyskuussa 2004 noin 5,9 miljoonaa paikannimeä. Nimen lisäksi tiedosto sisältää viittauksen paikkatietorekisteriin ja päivämäärän, jolloin nimeä koskevia tietoja on viimeksi muutettu. Koska tiedosto ei sisällä koordinaattitietoja, sen sovellusmahdollisuudet ovat varsin kapeat. Ladattava rekisteri on kuitenkin käyttökelpoinen esimerkiksi tunnistettaessa paikannimiä tekstistä [19].

ADL tarjoaa nettikäyttöliittymän hakujen tekemiseen paikkatietorekisteriin. Kuvassa 9 on esitelty käyttöliittymän asettelua. Haettaessa tietoja Joensuu-nimisestä paikasta, löytyi 23 kohdetta noin kahden sekunnin hakuajalla. Löytyneet kohteet näytetään linkkilistan lisäksi myös karttaikkunassa. Lisätietoja löytyneistä kohteista voi saada nopeasti valitsemalla haluttu paikannimi listasta. Tällöin valitusta kohteesta kerrotaan koordinaattitiedot sekä näytetään paikan sijainti tarkemmalla kartalla.



Kuva 9. Alexandria Digital Library:n Gazetteer-käyttöliittymä. Esimerkkihakuna Joensuu.

Eräs ADL Gazetteeriin hienouksista verrattuna muihin vastaaviin paikkatietorekistereihin on, että se sisältää myös paikkojen välisiä suhteita. Esimerkkeinä tällaisista suhteista ovat sisältyy-suhteet kuten on-osavaltiossa, on-valtiossa ja on-osa -suhteet [11]. Näin ollen rekisteri muodostaa hierarkian, joka sisältää hienojakoisempaakin sijainti-informaatiota kuin pelkän koordinaattitiedon.

ADL Gazetteerin toteutuksen päätarkoitus on ollut tehdä käyttöliittymä paikkatietorekisteriin. Tässä rekisteri onkin selvästi parempi kuin GNS käytössä olevista karttanäytöistä johtuen. Kuitenkin ADL:nkin käyttöliittymästä saa ahtaan vaikutelman, koska samalle näytölle on yritetty ahtaa liian paljon tietoa ja käytössä on pahimmillaan kolmekin vierityspalkkia yhtä aikaa. GNS puolestaan on parempi, jos käyttäjä tarvitsee omaan käyttöönsä paikkatietorekisteriä, joka sisältää paikannimien lisäksi muutakin kuin koordinaattitietoa.

3.3 Keltaiset sivut

Keltaiset sivut ovat alun perin olleet painetun puhelinluettelon osa, jossa yritykset ovat saaneet yhteys- ja tuotetietojaan näkyviin helposti saataville. Nykyisin keltaisten sivujen toimialue on laajentunut mobiilitekniikan alueelle siten, että palvelut ovat saatavina myös kännyköissä, kämmentietokoneissa ja internetissä. Keltaiset sivut toimivat kymmenissä eri maissa ja Suomessa palvelua hallinnoi *Suomen Keltaiset Sivut Oy* [21]. Painetun version eri osia julkaistiin Suomessa vuonna 2004 yhteensä 3,3 miljoonaa. Lisäksi www-palvelua käytti kuukausittain keskimäärin 200 000 käyttäjää tehden yhteensä noin miljoona hakua.

Tämän tutkielman lähtökohdista kiinnostavimmat Keltaisten sivujen valikoimasta löytyvät palvelut ovat *karttahaku* ja *reittihaku*. Karttahakua käytetään, kun halutaan etsiä sijainti tietylle osoitteelle. Osoite kirjoitetaan sille varattuun kenttään ja sovellus näyttää sen sijainnin kartalla. Kuvassa 10 on haettu osoitetta Jyväskylän keskustasta ja löytynyt osoite on merkitty kartan keskelle ympyrällä. Käyttöliittymä on helppokäyttöinen ja selkeä. Lisäksi käytettävissä on tulostimelle paremmin sovitettu kartta sekä lähestymiskartat, joiden avulla paikan löytäminen on helpompaa.

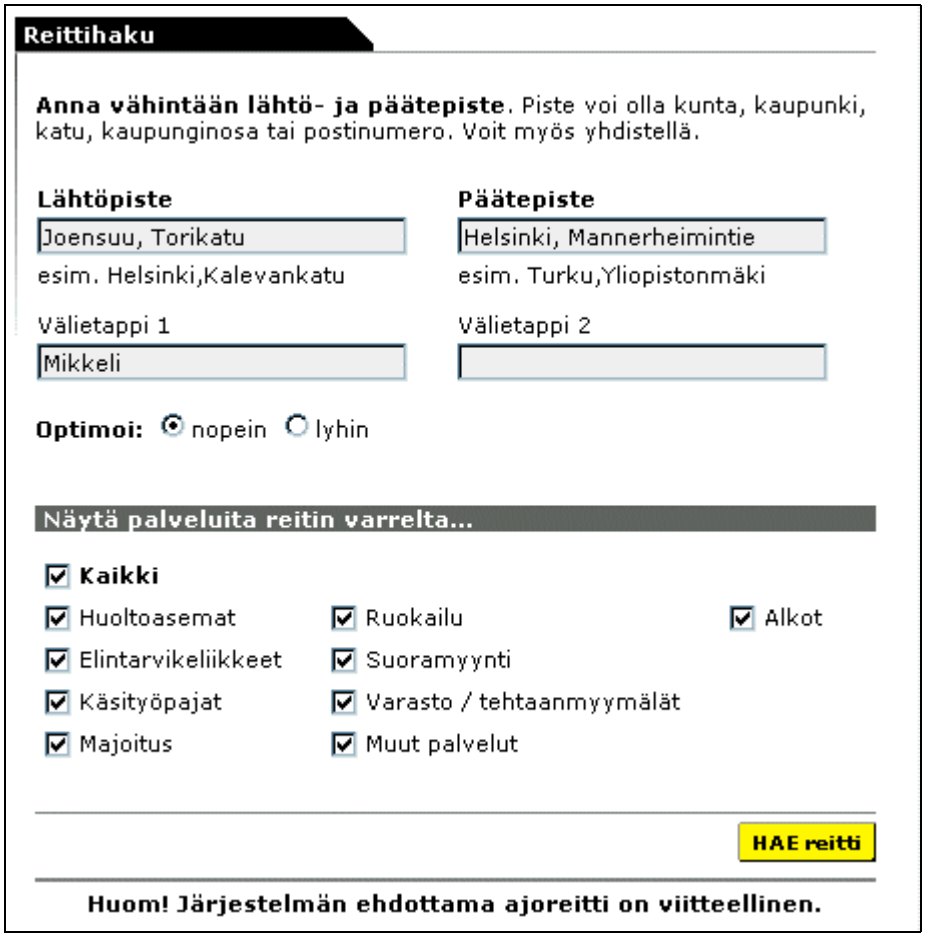
Kuva 10. Keltaisten sivujen karttahaun tulosikkuna.

Karttahaun ytimenä toimii osoiterekisteri, josta löytyy lähes kaikki Suomen postiosoitteet paikannustietoineen. Tarkkaa tietoa rekisterin laajuudesta ja peittävyydestä ei ole saatavilla, koska rekisteri sinällään ei ole julkinen vaan kyse on liiketoiminnasta. Osoiterekisterin lisäksi sovellus käyttää karttapohjia, joita löytyy koko Suomen kattavasta mittakaavasta 1:8000000 aina korttelitason mittakaavaan 1:5000 asti.

Karttahakua tehtäessä Keltaiset sivut etsittävän osoitteen lisäksi myös tietoja samasta karttanäkymästä löytyvistä muista palveluista. Lisäksi käyttäjä itse voi hakea toimialoittain tai hakusanoilla haluamiaan palveluita tarkastettavalta alueelta. Yritystietokannan kattavuudesta ei ole tarkkaa tietoa, mutta esimerkiksi Joensuun alueelta palvelu löytää neljä pizzeriaa. Kaupungista löytyy todellisuudessa ainakin 13 pizzeriaa. Yritysrekisterin neljästä pizzeriasta ainoastaan kaksi löytyy karttahausta. Joensuun 18:sta hotellista Keltaisten sivujen rekisteristä löytyy kuusi. Kaiken kaikkiaan yrityshaku toimii näppärästi, jos suuntaa matkansa tuntemattomalle paikkakunnalle ja

haluaa esimerkiksi tutkia palveluita, joita löytyy majoituspaikan läheltä, mutta sen kattavuus ei riitä kaikkien palveluiden löytymiseen.

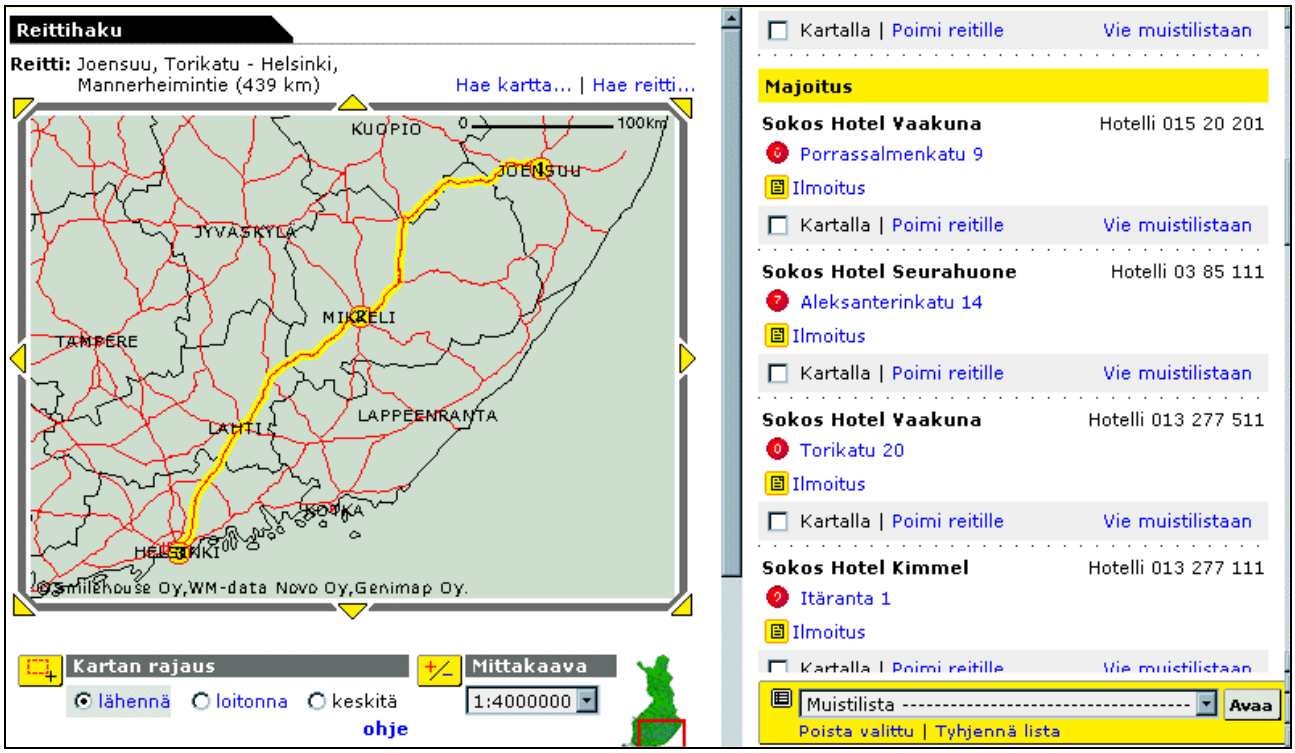
Reittihaku on karttahaun laajennus, jonka avulla voi etsiä reitin kahden eri osoitteen välillä. Lähtö- ja päätepisteen lisäksi reitille voi määritellä kaksi välietappia, joiden kautta matka halutaan kulkea. Kuvassa 11 on määritelty reitti Joensuusta Mikkelin kautta Helsinkiin. Matkan varrelta voidaan etsiä myös erilaisia palveluita, kuten huoltoasemia ja ruokailumahdollisuuksia. Lisäksi voidaan määritellä halutaanko etsiä nopein vai lyhin reitti. Reittihaun mahdollistaa sovelluksen käyttämä tierekisteri, johon on määritelty teitten numeroiden ja nimien lisäksi risteykset ja teillä käytössä olevat ajonopeudet. Kun tätä käytetään yhdessä osoiterekisteristä saatavien lähtö- ja päätepisteiden koordinaattien kanssa voidaan muodostaa reittejä pisteiden välille.



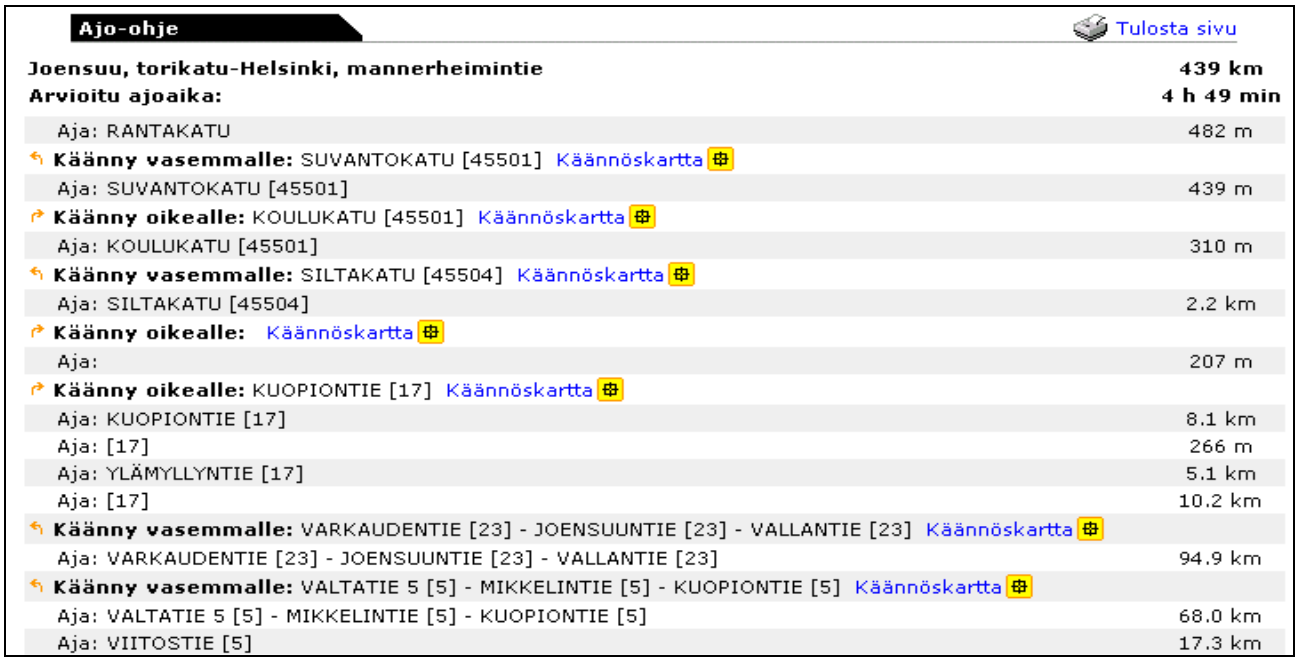
Kuva 11. Keltaisten sivujen reittihaun käyttöliittymä.

Käytettävyydeltään reittihaku on nopea ja helppo. Muodostetun ajoreitin voi näytöltä katseltavan version lisäksi tulostaa. Kuvassa 12 on reitti Joensuusta Mikkelin kautta Helsinkiin.

Karttamuotoisen tulosteen lisäksi reittihaku muodostaa myös tekstimuotoisen ajo-ohjeen. Esimerkki järjestelmän tuottamasta ajo-ohjeesta löytyy kuvasta 13. Ajo-ohje on viitteellinen, joskin se toimii varsin hyvänä tukena kartan kanssa käytettynä. Ajo-ohjeessa olevat tiennumerot tekevät oikeiden risteysten löytämisen helpoksi. Pienenä haittapuolena ajo-ohjeessa on se, että joissakin tapauksissa se opastaa käyttäjän lyhimmälle reitille vaikka karttamuotoinen ohje onkin muodostettu nopeinta reittiä käyttäen. Tämä ilmenee ristiriitaisuuksina näiden kahden ohjeen välillä. Lisäksi ongelmia aiheuttavat tilanteet, joissa tien nimi vaihtuu toiseksi, koska ajo-ohjeen lukija ei välttämättä tiedä, onko kysymyksessä risteys. Tällainen tilanne löytyy esimerkiksi tieltä numero 17, joka jossain vaiheessa muuttuu Ylämyllyntiestä Kuopiontieksi.



Kuva 12. Reittihaun tuottama kartta reitille Joensuu-Mikkeli-Helsinki.



Ajo-ohje Tulosta sivu

Joensuu, torikatu-Helsinki, mannerheimintie **439 km**
Arvioitu ajoaika: **4 h 49 min**

Aja: RANTAKATU	482 m
Käänny vasemmalle: SUVANTOKATU [45501] Käännöskartta	
Aja: SUVANTOKATU [45501]	439 m
Käänny oikealle: KOULUKATU [45501] Käännöskartta	
Aja: KOULUKATU [45501]	310 m
Käänny vasemmalle: SILTAKATU [45504] Käännöskartta	
Aja: SILTAKATU [45504]	2.2 km
Käänny oikealle: Käännöskartta	
Aja:	207 m
Käänny oikealle: KUOPIONTIE [17] Käännöskartta	
Aja: KUOPIONTIE [17]	8.1 km
Aja: [17]	266 m
Aja: YLÄMYLLYNTIE [17]	5.1 km
Aja: [17]	10.2 km
Käänny vasemmalle: VARKAUDENTIE [23] - JOENSUUNTIE [23] - VALLANTIE [23] Käännöskartta	
Aja: VARKAUDENTIE [23] - JOENSUUNTIE [23] - VALLANTIE [23]	94.9 km
Käänny vasemmalle: VALTATIE 5 [5] - MIKKELINTIE [5] - KUOPIONTIE [5] Käännöskartta	
Aja: VALTATIE 5 [5] - MIKKELINTIE [5] - KUOPIONTIE [5]	68.0 km
Aja: VIITOSTIE [5]	17.3 km

Kuva 13. Reittihaun tuottama ajo-ohje.

Yhteenvetona sanoisin, että Keltaisten sivujen tarjoamat kartta- ja reittipalvelut ovat riittävän tarkat ja monipuoliset tavallisen käyttäjän tarpeisiin. Käytössä olevat osoite- ja tierekisterit ovat hyviä esimerkkejä paikkatietorekistereiden kohdealueen mukaisesta soveltamisesta.

3.4 Google Local

Google Local on hakukone, joka käyttää hyväkseen paikkatietoa [6]. Perinteisesti hakija etsii verkosta tietoa eikä ole merkitystä mistä päin maailmaa haluttu tieto löytyy. Toisinaan kuitenkin halutaan tietoja palveluista, jotka ovat esimerkiksi kävelymatkan päässä tiedon etsijästä. Tällöin puhutaan *paikallisesta hausta* [4]. Google Local on *Google*-hakukoneen laajennus, jolla voi tehdä tällaisia hakuja.

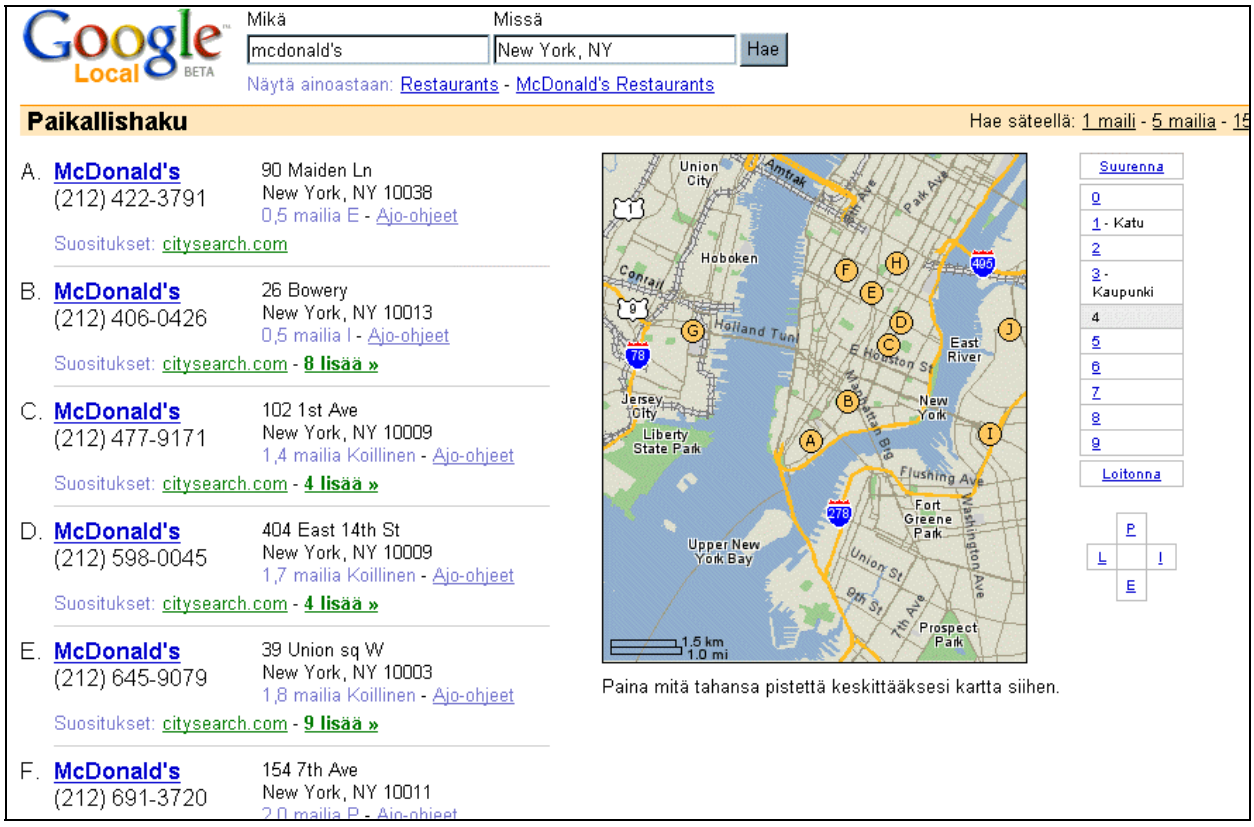
Google Local on julkaistu loppuvuodesta 2003 ja syyskuussa 2004 se oli yhä beta-testausvaiheessa, joskin vapaasti kaikkien käytettävissä. Google Local toimii toistaiseksi vain Yhdysvaltojen ja Kanadan palveluita haettaessa, mutta kehittäjillä on tarkoitus laajentaa se toimimaan myös muiden maiden paikkatiedoilla [6]. Paikallishaku toimii vain yrityksiä ja palveluita haettaessa, eikä sillä siis voi tehdä hakuja koko internetistä, vaan rajoitettua paikkatietorekisteriä hyödyntäen. Toisin sanoen Google Local ei ole alkuperäisen Googlen kaltainen koko webbiä indeksoiva hakukone vaan paikkatietorekisteriä hyödyntävä palvelu. Tästä lisää myöhemmin teknisen toteutuksen yhteydessä.

Kuvassa 14 on Google Local -hakukoneen käyttöliittymä. Hakua suoritettaessa on määriteltävä avainsanat sekä alue, jolta haku halutaan suorittaa. Sijainti voidaan antaa joko paikannimen tai postinumeron perusteella. Sijaintitiedon voi tallentaa siten, että hakukone ehdottaa oletuksena samaa sijaintia aina hakua tehtäessä. Sijainnin määrittämiseen kaipaisi kuitenkin jonkinlaista karttakäyttöliittymää, jolla käyttäjä voisi määrittää sijaintinsa tarkemmin. Jos käytetään suuren kaupungin nimeä kuten Los Angelesia sijaintitietona, ei ole ollenkaan varmaa, että hakutulokset tulevat siitä osasta kaupunkia, josta käyttäjä haluaisi.

Kuva 14. Google Local-hakukoneen käyttöliittymä.

Haun suorittaminen on todella nopeaa. Normaalin muutaman hakusanan sisältävän haun tulokset tulevat noin sekunnissa. Kuvassa 15 voi nähdä tulokset, jotka ovat tulleet etsittäessä New Yorkin McDonald's ravintoloita. Tulokset esitetään hakukoneen määrittelemässä tärkeysjärjestyksessä. Yrityksen nimen, puhelinnumeron ja osoitteen lisäksi näkyvillä on myös etäisyys ja suunta. Kuvan 15 tapauksessa välimatkat on mitattu jostain pisteestä Manhattanin eteläkärjen tuntumasta hakutuloksen B länsipuolelta, eikä käyttäjä pääse muuttamaan tätä pistettä.

Kuvassa 15 olevasta esimerkkitapauksesta voi huomata ongelman, joka aiheutuu sijainnin epätarkasta määrittämisestä. Koko kuvassa näkyvä alue kuuluu New Yorkiin ja on mahdollista, että käyttäjä ei haluakaan etsiä palveluita samasta osasta New Yorkia, johon Google Local olettaa haun kohdistuvan. Käyttäjä voi esimerkiksi etsiä palveluita Union Cityn kaupunginosasta, joka sijaitsee kuvassa 15 olevan kartan vasemmassa yläkulmassa. Tällöin löytyneet palvelut ovat liian kaukana halutuista tuloksista. Tämä ongelma pienenee, jos hakua tehdessä käytetään postinumeroita, jotka peittoalueeltaan ovat yleensä pienempiä kuin kaupungit.



Kuva 15. Esimerkki Google Local-hakukoneen tulossivusta.

Google Local-hakukoneen tekninen toteutus poikkeaa perinteisistä verkkohauista. Google Local ei ole koko verkon tietoja hyväksikäyttävä hakukone vaan lähinnä käyttöliittymä yhdysvaltain keltaisten sivujen (*Yellowpages*), tietokantoihin [6]. Tästä johtuen haut rajoittuvat vain palveluihin ja yritystietoihin. Tietokanta sisältää tiedon yrityksen osoitteesta ja käytössä on myös osoiterekisteri, jonka perusteella hakukone tietää yrityksen sijainnin myös koordinaattimuodossa.

Google Local käyttää hakutulosten rankkauksessa sekä palvelun etäisyyttä että sen omaa dokumenttien relevanttiuden arviointimenetelmä *PageRank*:ia. Tällainen lähestymistapa aiheuttaa epäloogisuutta esimerkiksi kuvassa 15 esillä olevassa haussa. McDonald's-ravintoloiden omasta verkkopalvelusta (www.mcdonalds.com) käy nimittäin ilmi, että eteläisellä Manhattanilla (hakutulosten A ja H välisellä alueella) on 19 McDonald's-ravintolaa, joista Google Local:in sadan ensimmäisen listalta löytyy 13. Google Local:in rankkausperusteilla esimerkiksi Greenwich Villagen McDonald's, joka sijaitsee hieman luoteeseen kuvan 15 tuloksesta B, on vasta 43. tärkein. Lyhyesti sanottuna Google Local ei välttämättä järjestä hakutuloksia parhaimmalla mahdollisella tavalla.

Paikkatiedon käyttämien hakuperusteena luo uuden kaupallisen lähestymistavan hakukoneen yhteyteen. Koska tiedetään tarkasti, mistä alueesta hakija on kiinnostunut, voidaan myös mainospaikkoja myydä siten, että ne näkyvät vain tietyllä alueella. Tällainen mainostaminen on järkevää silloin, jos kyseessä on esimerkiksi autokorjaamo tai vastaava pieni yritys, jonka asiakaskunta muodostuu lähiympäristön ihmisistä. Tällaisen yrityksen ei kannata markkinoida itseään maanlaajuisesti, vaan yritykset voivat ostaa hakusanapohjaisia mainospaikkoja. Mahdollisuus paikalliseen mainostamiseen on toteutettu myös perus-Googleen, jossa haun suorittajan paikantaminen suoritetaan joko hakusanojen tai käytössä olevan työaseman IP-osoitteen perusteella. Google Localin tarjoama paikallinen mainostus toimi syyskuussa 2004 vain Yhdysvaltojen alueella [6]. Mainostaja voi rajoittaa mainoksensa näkyvyyden osavaltioittain, kaupungeittain tai esimerkiksi 100 mailin säteellä Denveristä.

Google Local Beta vaikuttaa erittäin kehityskelpoiselta paikkatietoon perustuvalta hakukoneelta. Uskon, että se tulee laajenemaan myös muihin maihin ja paikallisuuteen perustuva mainostaminen tulee yleistymään. Paikallishaut sopivat parhaiten paljon liikkuville ihmisille ja omimmillaan ne ovat mobiiliviestimissä. Tällä hetkellä ongelma on, että Yellowpages:in käyttö hakujen perustana rajaa hakualuetta huomattavasti. Toinen vaihtoehtohan olisi erotella paikkatietoa web-dokumenteista esimerkiksi tässä tutkielmassa kuvatuilla tavoilla ja suorittaa yritysten paikannus sillä perusteella. Tällöin palvelu ei rajoittuisi ainoastaan kaupalliseen toimintaan.

4 Paikkatiedon eristämisen toteuttaminen

Kuten tässä tutkielmassa on aiemmin tullut esille, nettidokumenteista voidaan löytää eri muodoissa olevaa paikkatietoa. Löydettävissä olevan paikkatiedon määrästä ja laadusta on toistaiseksi tehty vain olettamuksia (luku 2) tai nojauduttu kirjallisuudessa esitettyihin arvioihin. Tässä luvussa esittelen toteuttamani ohjelman, jolla olen empiirisesti tutkinut suuntaa-antavalla tarkkuudella paikkatiedon esiintymistä nettidokumenteissa.

Tämä tutkielma edustaa omalla tavallaan alueensa perustutkimusta, sillä paikkatiedon ja webin yhdistämistä tutkielmassani esitellyillä tavoilla ei ole tutkittu kovinkaan laajasti, eikä varsinkaan valmiita sovelluksia löydy. Tutkimuksesta saatavat tulokset voivat antaa pohjaa muille sovelluksille, kuten esimerkiksi paikkatietoa hyödyntäville hakukoneille. Tulosten antama tieto kertoo onko paikkatietoa saatavilla tarpeeksi, jotta pidemmälle vietyjä sovelluksia olisi järkevää toteuttaa. Tämän tutkimuksen tulokset ovat kuitenkin vain suuntaa-antavia.

4.1 Hypoteesit

Osaltaan tulokset antavat tukea luvussa 2 esitellyille oletuksille paikkatietoa sisältävien elementtien esiintymistiheyksille. Tämän tutkimuksen laajuus ei kuitenkaan anna mahdollisuutta kaikkien paikkatietoa sisältävien elementtien eristämiseen, vaan rajaus on tehty helpoimmin ja yksikäsitteisimmin eroteltavissa oleviin elementteihin. Lisäksi tutkimustuloksissa otetaan kantaa, ovatko eri paikkatietoelementit toisistaan riippumattomia vai esiintyvätkö ne riippuvaisina ryhminä. Intuitiivisena oletuksena on, että esimerkiksi postinumero ja katunimi ovat toisistaan riippuvaisia.

Web-dokumentin koon ja siitä löytyneiden paikkatietoelementtien välillä voi olettaa olevan yhteyden. Mitä suurempi dokumentin koko sitä enemmän paikkatietoelementtejä siitä pitäisi löytyä. Tässä tutkimuksessa on tutkituista noin 24 000 web-dokumentista on määritelty, onko koolla yhteyttä löytyneiden paikkatietoelementtien määrään.

Verkon läpikäynti voidaan tehdä leveys- tai syvyyssuunnassa. Intuitiivisen oletuksen tekeminen siitä vaikuttaako läpikäynnin suunta löytyneiden paikkatietoelementtien lukumääriin, on vaikea tehdä. Tässä tutkimuksessa tehdään vertailu, onko läpikäynnin suunnalla vaikutusta löytyneisiin paikkatietoelementteihin.

WWW:n sisältö voidaan karkeasti jakaa kahteen sen perusteella, millainen sisältö dokumenteilla on. Suuri osa sisällöstä liittyy kaupallisiin palveluihin tai julkishallinnon verkkopalveluihin. Käytän näistä palveluista jatkossa nimitystä viralliset verkkopalvelut. Toisaalta merkittävä osa nettisivustoista on harrastelijoiden tekemiä ja ylläpitämiä epävirallisempia sivustoja. Kaupallisen toiminnan ja julkishallinnon luonteen perusteella etukäteisoletus on, että viralliset verkkopalvelut sisältävät enemmän paikkatietoa, kuin epäviralliset palvelut.

Web-dokumenteista eristettyjen paikkatietoelementtien perusteella voidaan tehdä päätelmiä, missä osiin Suomea verkkopalvelut painottuvat. Mikäli painotukset ovat samansuuntaisia asutuskeskittymien kanssa, voidaan olettaa että tiedon eristäminen on onnistunut varsin luotettavaksi.

4.2 Tutkimusasetelma

Tutkimus käsittää sekä eksaktia että epäsuoraa sijaintitietoa sisältävistä kohteista. Pääteltävissä olevia elementtejä, kuten yritys- ja henkilönimiä on rajattu tutkielman ulkopuolelle. Esimerkiksi henkilöiden ja yritysten nimien eristäminen ei olisi teknisesti kovinkaan hankalaa, mutta yksikäsitteisen sijaintitiedon liittäminen nimeen aiheuttaisi ongelmia ja tulosten tulkinta ei siten olisi välttämättä kovin luotettavaa. Osoitetiedon liittäminen nimitietoon esimerkiksi sähköisten puhelinluetteloiden kautta olisi taas liian kallista ja nimien yksiselitteisyys aiheuttaisi lisätarkastuksia.

Eksaktia sijaintietoa sisältävät geo-tagit ja osoite-elementit ovat mukana tutkimuskohteina. Osoite-elementistä kiinnostavin tieto on sen käyttöaste. Vaikka se kuuluukin HTML:n perustageihin, sen käyttöasteesta ei ole tarkkaa tietoa. Geo-tagit taas ovat vähemmän tunnettuja ja tietyllä tavalla verkon käyttäjille näkymättömiä, koska dokumentin ulkoasusta ei voi päätellä, onko niitä käytetty. Tutkimuksen tarkoituksena on tutkia, kuinka paljon niitä käytetään ja siten saada jotain käsitystä onko niillä nykyisellään käyttöarvoa.

Epäsuoraa paikkatietoa sisältäviä elementeistä tutkimuksessa ovat mukana tavallisimmat: paikannimet, katunimet, puhelinnumerot ja postinumerot. Kuitenkaan kaikkia mahdollisia nimiä ei ole etsitty. Tämä johtuu siitä, että tällöin olisi jouduttu käyttämään niin suuria nimirekistereitä, että parantuneiden tulosten saamiseksi olisi joutunut käyttämään huomattavasti enemmän aikaa. Tästä nimitietojen rajaamisesta johtuen tutkimus antaa vain suuntaa-antavia tuloksia.

Päätin myös rajata tutkimusalueen koskemaan vain suomalaisia WWW-sivustoja. Pohjois-Amerikassa vastaavantyyppisiä tutkimuksia on tehty aiemminkin tunnetuin tuloksin [13]. Koska suomen kieli poikkeaa kieliopillisesti selvästi anglo-saksisista kielistä, on nimien kerääminen tekstistä hieman vaikeampaa. Esimerkiksi englannin kielessä paikannimet esiintyvät lähes poikkeuksetta perusmuodossaan, jolloin niiden erottelu on huomattavasti helpompaa kuin suomen kielessä, jossa sanat voivat esiintyä useilla tavoilla taivutettuina.

Käytännössä tein tutkimusalueen rajauksen siten, että tutkin ainoastaan fi-päätteisiä sivustoja. Rajauksen toteuttaminen on helppoa, koska tällöin dokumenttien sisältämille teksteistä ei tarvitse erikseen tutkia millä kielellä ne on kirjoitettu. Tutkittua tietoa siitä kuinka suuri osuus fi-päätteisistä sivustoista on suomenkielisiä ei ole saatavilla. Perusoletus kuitenkin on, että suurin osa tutkituista sivuista on suomenkielisiä.

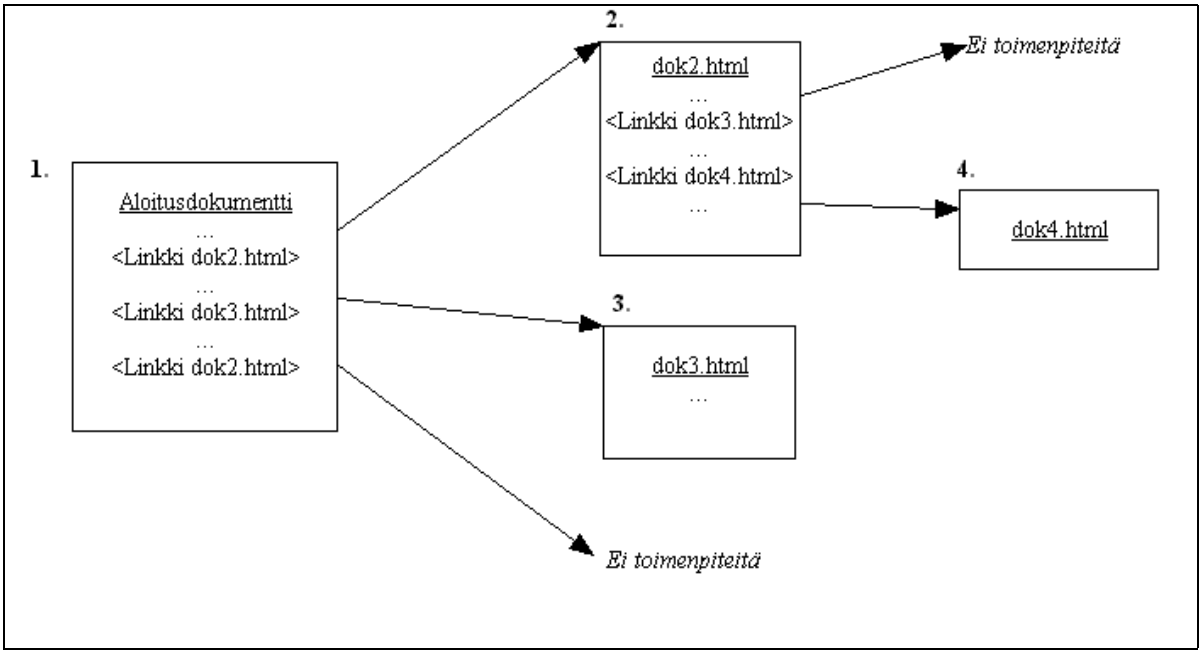
Fi-päätteisiä internet-osoitteita hallinnoi Suomen Viestintävirasto [23]. Domainien myöntämisperusteet helpottuivat 1.9.2003, jonka jälkeen fi-päätteisen osoitteen on voinut hankkia melkein kuka tahansa. Aiemmin kyseisiä osoitteita myönnettiin vain yrityksille, tuotemerkeille ja muille vastaaville vakiintuneille nimikkeille. Syyskuussa 2004 suomalaisia domaineja oli myönnetty yli 86000. K- ja s-kirjaimilla alkavia oli eniten, kumpaakin noin 8000 kappaletta.

Sivustojen läpikäynti on suoritettu käytännössä useammissa pienemmissä, noin 10000 sivua kattavissa ajoissa. Läpikäyntityökalu on selainkäyttöinen, eikä sen vakaus ole riittävä useita vuorokausia kestäviin suuriin ajoihin. Tutkimuksessa käytettyä sovellusta esitellään tarkemmin myöhemmin.

Varsinainen sivujen läpikäynti suoritetaan leveyshaulla kuvan 16 esimerkin mukaisesti siten, että läpikäynti aloitetaan aloitusdokumentista, josta etsitään kaikki mahdolliset linkit muihin dokumentteihin. Ensimmäisenä löytyvät viittaukset osoittavat *dok2.html* ja *dok3.html* -nimisiin dokumentteihin. Löytymisjärjestyksestä määräytyy dokumenttien läpikäyntijärjestys. Dok2.html-nimiseen dokumenttiin löytyy vielä toinenkin viittaus, mutta se ei aiheuta toimenpiteitä, koska jokainen dokumentit tutkitaan vain kerran.

Kun aloitusdokumentti on tutkittu loppuun asti, otetaan käsittelyyn dok2.html. Ensimmäisenä siitä löytyy viittaus dok3.html-dokumenttiin, joka ei aiheuta toimenpiteitä, koska vastaava viittaus on löydetty jo aiemmin aloitusdokumentista. Tämän jälkeen löydetään vielä viittaus dok4.html-

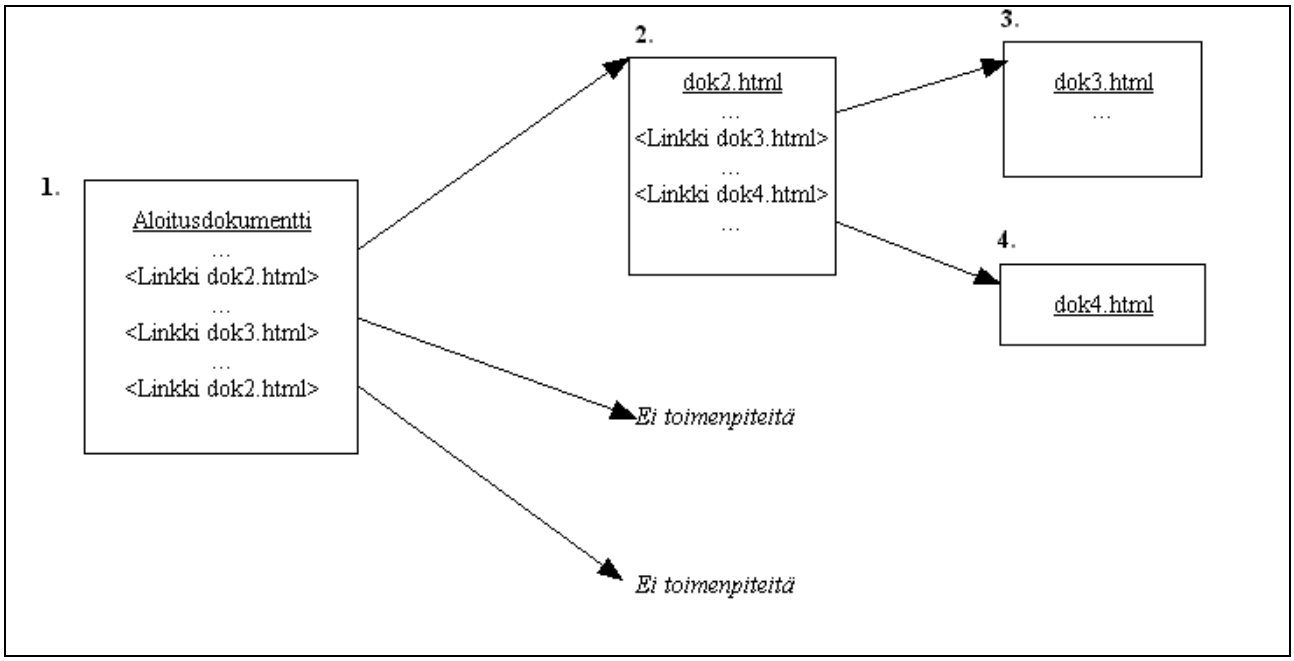
dokumenttiin. Lopuksi suoritetaan läpikäynnit dok3.html ja dok4.html-dokumentteihin, minkä jälkeen kaikki esimerkin sisältämät dokumentit on tutkittu.



Kuva 16. Esimerkki leveyssuuntaisesta läpikäynnistä.

Leveyssuuntaista läpikäyntiä on tässä tutkimuksessa rajattu siten, että kaikista sivustoista tutkitaan vain etusivu, eli alisivustoihin ei paneuduta. Tämä tarkoittaa sitä, että paljon olemassa olevaa paikkatietoa saattaa jäädä löytymättä. Esimerkiksi joillain sivustoilla ensimmäinen sivu on toteutettu siten, että se tutkii käytössä olevan selaimen ja suorittaa ohjauksen juuri sille optimoiduille sivuille. Tällaisissa tapauksissa varsinainen kohdesivu jää löytymättä.

Toinen vaihtoehto olisi tutkia sivustoja syvyyssuunnassa. Kuvassa 17 on esitetty esimerkki, kuinka dokumenttien läpikäynti tapahtuu syvyyssuunnassa. Tälläkin kertaa aloitusdokumentista etsitään ensin kaikki viittaukset muihin dokumentteihin. Jälkimmäinen viittaus dok2.html-nimiseen dokumenttiin jää tälläkin kertaa huomiotta. Syvyyssuuntaisen läpikäynnin periaatteen mukaisesti seuraavaksi etsitään dok2.html-dokumentista kaikki linkit ja seurataan niitä niin syvälle kuin mahdollista. Tässä tapauksessa ainoat löytyvät viittaukset ovat dokumentteihin dok3.html ja dok4.html. Koska kaikki dok2.html-dokumentista löytyneet linkit on tutkittu, palataan takaisin aloitusdokumentista löytyneisiin viittauksiin, joita tässä tapauksessa on jäljellä vain yksi. Kyseessä on viittaus dok3.html-dokumenttiin, mutta koska kyseinen dokumentti on jo tutkittu, voidaan tämäkin linkki jättää huomiotta. Syvyyssuuntainen haku on siis luonteeltaan rekursiivista.



Kuva 17. Esimerkki syvyyssuuntaisesta läpikäynnistä.

Läpikäynti syvyyssuunnassa voitaisiin toteuttaa ainakin kahdella eri tavalla. Sivuston läpikäynti päättyisi silloin, kun löydetään paikkatietoa tai kun kaikki sivustolta löytyvät dokumentit on käyty läpi. On selvää, että syvyyssuuntainen läpikäynti lisäisi merkittävästi tutkittavien dokumenttien määrää. Nykyisten tietokantapohjaisten internet-palveluiden aikakaudella ei ole ollenkaan epätavallista, että yksi ainoa sivusto voi sisältää tuhansia sivuja. Voi olla, että kaikilla saman sivuston sivuilla on samanlaisena toistuvia osia, esimerkiksi yrityksen osoite ja puhelinnumero alamarginaalissa Tämän takia tutkimustulokset saattaisivat vääristyä yksittäisten sivustojen osuuden painottuessa liikaa. Kaiken kaikkiaan kaikkien fi-päätteisten sivustojen kaikkien sivujen läpikäyminen tämän tutkielman asettamissa rajoissa ei ole realistista.

Leveyssuuntainen läpikäynti mahdollistaa varsin kattavan läpikäynnin suomalaisiin sivustoihin. Koska lähes kaikki mahdolliset sivustot ovat osana tutkimusta, on tuloksena jonkinlainen poikkileikkaus koko fi-päädomainista.

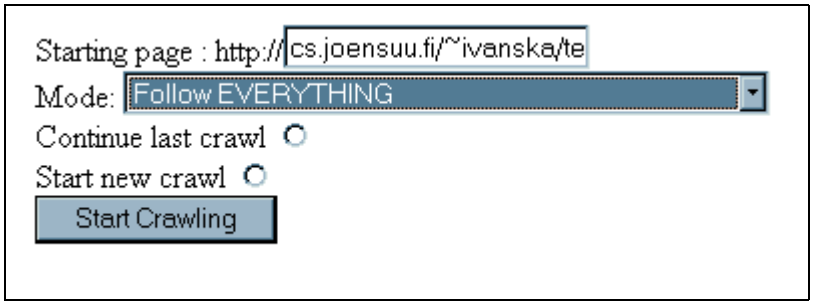
4.3 Sovellus

Varsinaisen sovelluksen, joka suorittaa WWW-dokumenttien läpikäynnin ja etsii niistä paikkatietoa sisältäviä elementtejä, olen toteuttanut PHP-kielellä. Sovellusta käytetään internet-selaimella ja sen ydin on *crawler*, joka suorittaa sivujen läpikäynnin.

Crawler on käyttöliittymältään mahdollisimman yksinkertainen. Sille määrätään etusivu, josta se aloittaa verkon tutkimisen. Lisäksi voidaan määrätä hakumoodi, jolla hakua voidaan rajata. Crawler voidaan asettaa jollain seuraavista tavoista:

- Seuraamaan kaikkia löytämiään viittauksia muihin dokumentteihin.
- Seuraamaan vain saman domainin alaisia linkkejä (esimerkiksi *joensuu.fi* sisältää *cs.joensuu.fi* ja *www.joensuu.fi*).
- Seuramaan vain täsmälleen saman isännän alaisia linkkejä (esimerkiksi vain *cs.joensuu.fi*).
- Seuraamaan täsmälleen saman isännän ja polun alaisia linkkejä (esimerkiksi vain *cs.joensuu.fi/~ivanska/*).
- Olemaan seuraamatta mitään linkkejä. Tällöin vain määritellystä aloitussivusta etsitään sijainti-informaatiota.

Kuvassa 18 crawler on käyttövalmiina. Verkon läpikäynti aloitetaan *Starting page* -kentässä määritellyltä sivulta ja kaikkia löydettyjä linkkejä seurataan eteenpäin. Kuvassa näkyvät valintapainikkeet edellisen läpikäynnin jatkamiseksi tai uuden aloittamiseksi eivät ole käytössä.



Kuva 18. Crawlerin käyttöliittymä.

Crawler selaa verkkoa aiemmin esitellyn leveyshaun mukaisesti. Etsiessään paikkatietoja dokumenteista sovellus tulostaa näytölle tietoja läpikäynnin etenemisestä. Kuvassa 19 crawler on tutkimassa d-kirjaimella alkavia fi-päätteisiä sivustoja. Tulosteesta käy ilmi, kuinka monta sivustoa crawler on kyseisessä ajossa tutkinut ja milloin läpikäynti on suoritettu. Näytölle tulostetuista linkeistä käyttäjä voi ajon aikana katsoa millaisia tutkittavat dokumentit ovat. Mikäli dokumentista on löytynyt tunnettuja paikannimiä, näiden pituus- ja leveyskoordinaattien minimi- ja maksimiarvot tulostetaan linkin alapuolelle.

Kuva 19. Crawler toiminnassa.

Sivustojen läpikäynti on dokumentoitu ja kaikki yksittäiset tulokset on tallennettu. Tutkimustuloksia selvitettäessä on käytetty tätä dataa, josta jatkossa käytän nimitystä *tulosjoukko*. Tulosjoukko on XML-muodossa, jolloin sen siirrettävyys eri ympäristöihin on helppoa ja tutkimustulosten laskeminen on vaivatonta.

4.4 Sovelluksen suorituskyky

Sovelluksen testaamiseksi suoritin sille testiajon, jossa otosjoukkona oli sata fi-päätteistä sivustoa. Tutkin näiden sisältämät paikkatietoelementit sekä manuaalisesti että sovelluksen avulla. Näitä tuloksia vertailemalla saa kohtalaisen käsityksen siitä, kuinka paljon sovelluksesta johtuvia virheitä tutkimustulokset sisältävät.

Taulukossa 1 on esitelty testauksen tuloksia. Manuaalisen tarkastuksen sarakkeessa ovat niin sanotut oikeat tulokset, jotka käsin suoritetussa tarkastelussa on löydetty. Sovelluksen suorittamassa läpikäynnissä on neljä saraketta, joissa ovat tiedot automaattisesta läpikäynnistä. *Yhteensä*-sarake kertoo kaikki sovelluksen löytämien elementtien lukumäärät. *Oikeat*-sarake kertoo, kuinka monta löydettyä elementtiä on ollut samoja kuin manuaalisessa tarkastelussa. *Väärät*-sarakkeessa on tieto

siitä, kuinka moni sovelluksen löytämistä elementeistä ei ole sisältänyt sitä informaatiota, mitä olisi pitänyt. Esimerkiksi postinumeroksi tulkittu numerosarja onkin tarkoittanut jotain muuta. Elementit, joita sovellukselta on jäänyt löytämättä on summattu *puuttuvat*-sarakkeseen.

Recall-arvo kertoo kuinka suuren osan kaikista mahdollisista elementeistä sovellus on löytänyt [14]. Esimerkiksi puhelinnumeroista sovellus on löytänyt 35 / 45 = 78 % kaikista mahdollisesta. *Precision* taas kuvaa suoritustarkkuutta. Se kertoo kuinka suuri osa kaikista sovelluksen löytämistä elementeistä on ollut oikeita [14]. Esimerkiksi löydetyistä 24:stä postinumerosta 22 on ollut oikeita, jolloin precision on 22 / 24 = 92 %.

Taulukko 1. Käytössä olleen sovelluksen virhealttius.

	Manuaalinen tarkastus	Sovelluksen suorittama läpikäynti					
		Yhteensä	Oikeat	Väärät	Puuttuvat	Recall	Precision
Sivustoja, joilla paikkatietoa	31	31	28	3	3	0,90	0,90
Löytyneet paikannimet	120	75	75	0	45	0,63	1,00
Löytyneet puhelinnumerot	45	56	35	21	10	0,78	0,63
Löytyneet kadunnimet	27	42	27	15	0	1,00	0,64
Löytyneet postinumerot	22	24	22	2	0	1,00	0,92

Yhteenvetona voidaan todeta, että sovelluksen suorituskyvyssä olisi parantamisen varaa. Kaikki paikannimet jotka sovellus löytää, ovat oikeita, mutta liian suuri osa niistä jää löytymättä. Tämä johtuu siitä, että käytettävässä nimirekisterissä ei ole kuin 300 paikannimeä, jolloin kaikki muut nimet jäävät löytymättä. Tällaisia ovat esimerkiksi pienet paikkakunnat ja tunnetut matkailukohteet. Lisäksi paikannimistä on hyväksytty vain ne, jotka on kirjoitettu isolla alkukirjaimella. Sovellus ei myöskään löydä nimiä, jotka eivät esiinny yksinään. Esimerkiksi Itä-Helsinkiä sovellus ei tunnista Helsingiksi.

Puhelinnumeroiden tunnistuksessa suurimmat ongelmat ovat puhelinnumeroiden kirjoitusasujen suuri vaihtelevuus sekä numeeriset merkkijonot, jotka tarkoittavat jotakin muuta kuin puhelinnumeroa. Nämä aiheuttavat sen, että löytyy paljon vääriä puhelinnumeroita ja toisaalta paljon numeroita jää myös puuttumaan.

Kadunnimien suhteen testitulos on recallin osalta yllättävänkin hyvä. Koska sovellus tunnistaa kadunnimiksi vain *tie*-, *katu*-, *polku*- ja *kuja*-päätteiset sanat, on hieman yllättävää ettei esimerkiksi *raitti*-päätteisiä kadunnimiä ollut ainoatakaan. Tositilanteessahan näitä on jäänyt huomaamatta.

Precision on liian huono käytetystä tutkimusmenetelmästä johtuen. Muun muassa sanat *laukkuja* ja *seurakuntien* ovat tulleet tunnistetuiksi kadunnimiksi suotuisista loppuosistaan johtuen.

Postinumeroiksi on luokiteltu kaikki merkkijonot, jotka koostuvat viidestä merkistä. Recall-luvusta tulee näin ollen helposti korkea, koska Suomessa postinumerot ovat aina viisinumeroisia. Precisionia taas laskee se, että hyväksytyiksi tulevat kaikki viisinumeroiset merkkijonot riippumatta niiden tarjoamasta todenmukaisuudesta.

Address-tagin ja *geo-tagien* etsiminen on helpompaa kuin edellä mainittujen elementtien. Address-elementeiksi sovellus hyväksyy kaikki löytyneet *<ADDRESS>* - *</ADDRESS>*-parit ja tallentaa niiden väliin jäävän tekstin. Geo-tagien tunnistaminen on suoritettu etsien tekstistä merkkijonoja *geo.position*, *geo.placename* ja *geo.region*.

4.5 Tulokset

Tutkimuksessa tutkittiin paikkatiedon esiintymistä kuudesta eri näkökulmasta:

- paikkatiedon esiintymistiheys web-dokumenteissa
- paikkatietoelementtien esiintymisen riippumattomuus toisista paikkatietoelementeistä
- web-dokumentin koon vaikutus löytyneisiin paikkatietoelementteihin
- paikkatiedon esiintyminen harrastelijoiden ylläpitämillä sivustoilla
- syvyys- ja levyssuuntaisten verkon läpikäyntien väliset erot
- löytyneiden paikkatietoelementtien painottuminen Suomen eri osien välillä

4.5.1 Paikkatietoelementtien esiintymistiheydet web-dokumenteissa

Esiintymistiheyksiä tutkittaessa kohteena olivat sivustot, joiden www-osoite päättyy fi-päätteeseen. Fi-päätteisiä domaineja on myönnetty tutkimuksen ajankohtaan mennessä noin 86000 kappaletta. Näistä noin 20000 ei ole varsinaisesti käytössä, vaan ne on vain varattu mahdollista myöhempää käyttöä varten. Lisäksi sovellus on karsinut pois kaikki ne sivustot, joiden palvelin on läpikäynnin aikana palauttanut jonkin muun kuin OK-statuksen. Näiden karsimistoimenpiteiden jälkeen läpikäynti on suoritettu noin 24000 sivustolle.

Taulukossa 2 on vertailtu tutkittuja sivustoja niiden koon perusteella. Taulukosta käy ilmi, että tyypillisesti tutkitut sivustot ovat kooltaan varsin pieniä, yleensä alle 20 kilotavua. Lähes puolet

tutkituista dokumenteista on kooltaan alle 2 kilotavua. Sivujen keskikoko on hieman alle 5000 merkkiä. Koska HTML-kieli perustuu tageihin, joita välttämättä on jokaisella sivulla, on selvää että varsinaista leipätekstiä ei pienillä sivuilla ole kovinkaan paljon.

Taulukko 2. Tutkitut sivustot koon mukaan lajiteltuna.

Sivuston koko	Lukumäärä	Osuus
0-500t	3364	14 %
500-1000t	4808	20 %
1-2kt	3916	16 %
2-5kt	5182	21 %
5-10kt	3969	16 %
10-20kt	2164	9 %
20-50kt	749	3 %
50-100kt	90	0,4 %
yli 100kt	19	0,1 %
Sivustoja yhteensä	24261	100 %

Sivustojen läpikäynnin suorittaneen sovelluksen tarkkuudessa on parantamisen varaa. Testausvaiheessa ilmeni, että tuloksiin tulee mukaan sellaisiakin elementtejä, jotka tulkitaan paikkatietoelementeiksi, vaikka ne eivät sisälläkään haluttua informaatiota. Tästä johtuen suoritin tuloksille vielä manuaalisen tarkastuksen, jossa poistin tuloksista ne elementit, joista ilmenee ilman asiayhteyttäkin, etteivät ne sisällä paikkatietoinformaatiota. Tällaisia tapauksia ovat muun muassa kadunnimiksi tulkitut merkkijonot, jotka selvästikään eivät ole katunimiä, kuten sana seurakuntien. Lisäksi tarkistin postin verkkopalvelun postinumeroluettelosta kaikki postinumerot ja eliminoin postinumeroista kaikki viisinumeroiset luvut, jotka eivät ole postinumeroita [22]. Merkittävän vääristymän tuloksissa aiheutti myös numerosarja 21600, jonka voi tulkita myös Paraisten postinumeroksi. Kyseinen numero generoituu dokumenttiin useita kertoja sisältäen jotakin muotoilutietoa mikäli HTML-editorina käytetään Microsoft Wordia. Puhelinnumeroille tein silmämääräisen tarkistuksen, jossa poistin selvästi mahdottomat tapaukset. Address-tageista poistin ne, jotka eivät sisältäneet paikkatietoa. Seuraavissa taulukoissa esittelen tulokset sekä ennen manuaalista korjausta että korjauksen jälkeen. Korjauksen jälkeiset tulokset ovat siis lähempänä todellista tilannetta kuin ennen korjausta olevat.

Tarkistuksen jälkeen löydettyjä paikkatietoelementtejä jäi jäljelle yli 40000 kappaletta. Sovellus tunnisti teksteistä alkujaan lähes 49000 elementtiä, joista virheellisiksi osoittautui ainakin 9000 kappaletta. Kuten jo testausvaiheessa ilmeni paikannimiä ja puhelinnumeroita dokumenteissa mitä ilmeisimmin olisi huomattavasti enemmänkin, kun taas katunimistä ja postinumeroista lähes kaikki mahdolliset sisältynevät korjattuun lukumäärään. Taulukossa 3 on esitelty löydettyjen paikkatietoelementtien esiintymislukumääriä.

Taulukko 3. Löydettyjen paikkatietoelementtien lukumäärät.

	Kaikki	Validit
Paikannimiä	21323	100 %
Kadunnimiä	7567	74 %
Postinumeroita	10044	49 %
Puhelinnumeroita	9870	84 %
Osoite-elementtejä	159	4 %
Yhteensä	48963	82 %

Vähintään yksi paikkatietoelementti löytyi 35 prosentista tutkituista sivustoista. Eri muodoissaan esiintyneet paikannimet olivat selkeästi yleisin paikkatiedon esitystapa. Taulukossa 4 on esitetty, kuinka suuri osa tutkituista dokumenteista sisälsi eri paikkatietoelementtejä.

Taulukko 4. Paikkatietoelementtejä sisältävien dokumenttien suhteellinen osuus kaikista tutkituista dokumenteista

Elementtityyppi	Lkm	Osuus
Paikannimi	6547	27,0 %
Katunimi	4031	16,6 %
Postinumero	3992	16,5 %
Puhelinnumero	3621	14,9 %
Vähintään yksi	8486	35,0 %

Dokumentit sisältävät toisistaan poikkeavia määriä paikkatietoelementtejä. Mitä useampia elementtejä voidaan löytää, sitä luotettavammin dokumentti voidaan kohdistaa jollekin

maantieteelliselle alueelle. Taulukossa 5 on tarkemmin esitelty niiden sivustojen lukumääriä, joista on löytynyt paikkatietoelementtejä.

Taulukko 5. Löydettyjen validien paikkatietoelementtien lukumäärät sivuittain esitettynä.

Elementit	Validit	Osuus
1	1989	8,2 %
2	1391	5,7 %
3	1350	5,6 %
4	1053	4,3 %
5	852	3,5 %
6-7	812	3,3 %
8-10	489	2,0 %
11-15	297	1,2 %
16-20	111	0,5 %
21-50	106	0,4 %
51-100	19	0,1 %
yli 100	15	0,1 %
Yhteensä	8486	35,0 %

Yhdestä dokumenteista löytyi enimmillään 432 paikkatietoelementtiä. Keskimäärin yhdellä tutkitulla sivulla oli 1,7 hyväksyttyä paikkatietoelementtiä. Mikäli lukuun otetaan mukaan vain ne dokumentit, joissa on vähintään yksi paikkatietoelementti, on vastaava luku 4,9. Toisin sanoen niillä sivuilla, joilla on paikkatietoa sisältäviä elementtejä, on niitä yleensä enemmän kuin yksi.

Epäsuoraa paikkatietoa sisältäneiden elementtien lisäksi tutkin myös eksaktia paikkatietoa sisältävien osoite-elementtien (address-tagien) ja *geo-tagien* esiintymistiheyttä. Osoite-elementtien esiintyminen oli yllättävänkin harvinaista. Yhtensä 24261:n sivun joukosta kyseisiä tageja löytyi vain 133 kappaletta, joista ainoastaan 6 sisälsi tunnistettavan osoitetiedon tai puhelinnumeron. Ylijäävät runsaat sata löydettyä osoite-elementtiä olivat lähes poikkeuksetta *Apache*-WWW-palvelimen tuottamia ilmoituksia, joissa se ilmaisi käytetyn palvelimen tyypin. Esimerkkinä tällaisesta voisi olla seuraava ilmoitus:

```
<ADDRESS>Apache/1.3.26 Server at www.laajasalonkokoomus.fi Port 80</ADDRESS>
```

Geo-tagien käyttöastetta tutkin erillisenä ajona, jossa kävin läpi 23200 fi-päätteistä sivustoa samalla periaatteella kuin muitakin paikkatietoelementtejä. Kyseisiä tageja ei valitusta otoksesta löytynyt ainoatakaan. Tämäkin tulos olisi mitä ilmeisimmin ollut erilainen, mikäli tutkimus olisi pureutunut syvemmälle verkon uumeniin ja harrastajien verkkopalveluihin. Tuloksesta voi kuitenkin päätellä, ettei kyseinen menetelmä ole kovinkaan yleinen eikä laajassa käytössä.

Varsinaisen sisältönsä lisäksi myös web-dokumentin osoite voi sisältää paikkatietoelementtejä. Tyypillisiä ovat esimerkiksi kuntien palvelut, kuten *www.ulvila.fi* tai yritysten palvelut kuten *www.helsinginautohuolto.fi*. Tutkin tällaisia tilanteita siten, että etsin tunnistettavia paikkatietoelementtejä tutkimustuloksissa olleiden dokumenttien osoitteiden aluista. Tällöin esimerkiksi *www.enonkuljetus.fi* tuli hyväksytyksi, kun taas *www.hienohomma.fi* ei, vaikka molemmissa löydettävissä Eno-nimisen kunnan nimi. Tällä tavalla tehdyn läpikäynnin tuloksena löysin 1522 paikannimeä tutkituista 24261 sivuista, mikä on 6,3 % kaikista osoitteista. Lisäksi www-osoitteista löytyi 10 kadunnimeä.

Tutkituista dokumenteista löytyneen paikkatiedon määrä on mielestäni riittävä esimerkiksi paikallishakukoneen tarpeisiin. World Wide Web:issä on tämän tutkimuksen perusteella oltava ainakin useita satoja tuhansia ellei miljoonia suomenkielisiä dokumentteja, jotka sisältävät paikkatietoa.

4.5.2 Paikkatietoelementtien riippumattomuus

Paikkatietoelementtien riippumattomuutta tutkimalla voidaan tehdä päätelmiä, ovatko löydetyt elementit todella sisältäneet paikkatietoa vai ovatko esimerkiksi postinumeroiksi tulkitut numerosarjat sisältäneet jotakin muuta informaatiota. Mikäli elementit ovat esiintyneet toisistaan riippumattomasti, ne mitä ilmeisimmin eivät ole sisällöltään paikkatietoa. Tällaisen päätelmän voi tehdä esimerkiksi siitä, että katunimet ja postinumerot esiintyvät usein yhdessä, koska molemmat ovat olennaisia osia postiosoitteesta, jota ei voi eksaktisti esittää, jos jompikumpi puuttuu. Samanlainen tilanne on myös paikannimien ja postinumeroiden yhteisessä esiintymisessä.

Paikkatietoelementtien riippumattomuuksia arvioitaessa ensimmäinen vaihe on laskea yksittäisten elementtien esiintymistodennäköisyydet dokumentissa. Taulukossa 6 nämä todennäköisyydet on laskettu käytössä olleesta materiaalista. Esimerkiksi todennäköisyys sille, että dokumentti sisältää katunimen on 17 % ja sille, että dokumentti ei sisällä katunimeä on 83 %.

Taulukko 6. Paikkatietoelementtien esiintymistodennäköisyydet dokumenteissa.

Elementti	On	Ei	P(”On”)	P(”Ei”)
Paikannimi	6547	17714	27 %	73 %
Katunimi	4031	20230	17 %	83 %
Postinumero	3992	20269	16 %	84 %
Puhelinnro	3621	20640	15 %	85 %

Mikäli elementit ovat toisistaan riippumattomia, yhdistetty todennäköisyys kahden tai useamman elementin esiintymiselle voidaan laskea todennäköisyyksien tulona [20]. Esimerkiksi todennäköisyys sille, että dokumentti sisältää sekä paikannimen että kadunnimen tulisi olla:

$$P(paikannimi \cap kadunnimi) = P(paikannimi) * P(kadunnimi) = 0{,}27 * 0{,}17 = 0{,}05$$

Taulukossa 7 on laskettu todennäköisyydet kaikille paikkatietoelementtien esiintymiskombinaatioille käyttäen taulukon 6 arvoja yksittäisten esiintymien todennäköisyyksinä. Plus-merkki sarakkeessa tarkoittaa, että dokumentti sisältää elementin ja miinus-merkki sitä, että dokumentti ei sisällä elementtiä. Laskennalliset todennäköisyydet on laskettu yllä esitetyn riippumattomien elementtien todennäköisyyskaavan avulla ja todelliset todennäköisyydet edustavat kokeellisten testitulosten mukaisia todennäköisyyksiä.

Taulukko 7. Paikkatietoelementtien kombinaatioiden esiintymistodennäköisyydet dokumenteissa.

Selite	Paikannimi	Katunimi	Postinro	Puh.nro	Lkm	Laskenn.	Todell.
Kaikki	+	+	+	+	1550	0,1 %	6,4 %
	+	+	+	-	759	0,6 %	3,1 %
	+	+	-	+	204	0,6 %	0,8 %
	+	+	-	-	361	3,2 %	1,5 %
	+	-	+	+	296	0,6 %	1,2 %
	+	-	+	-	183	3,1 %	0,8 %
	+	-	-	+	277	2,8 %	1,1 %
Paikannimi	+	-	-	-	2917	16,0 %	12,0 %
	-	+	+	+	580	0,3 %	2,4 %
	-	+	+	-	274	1,7 %	1,1 %
	-	+	-	+	109	1,5 %	0,4 %
Kadunnimi	-	+	-	-	194	8,6 %	0,8 %
	-	-	+	+	173	1,5 %	0,7 %
Postinro	-	-	+	-	177	8,5 %	0,7 %
Puh.nro	-	-	-	+	432	7,6 %	1,8 %
Ei mitään	-	-	-	-	15775	43,3 %	65,0 %
Yhteensä					24261	100 %	100 %

Taulukosta 7 voidaan huomata, että tutkittavista sivustoista lasketut paikkatietoelementtien esiintymistodennäköisyydet poikkeavat usein merkittävästi laskennallisista todennäköisyyksistä. Tilanteissa, joissa dokumentista löytyy vähintään kolme neljästä tutkitusta elementistä, todellinen esiintymistodennäköisyys on miltei poikkeuksetta merkittävästi suurempi kuin laskennallinen todennäköisyys. Perusteluna tälle ilmiölle on mielestäni se, että yhteystietoja julkaistaessa halutaan tuoda ilmi sekä postiosoite että puhelinnumero.

Vastaavasti laskennallisia todennäköisyyksiä selvästi harvinaisempia ovat tilanteet, joissa yksittäinen paikkatietoelementti esiintyy yksinään. Kadunnimet ja postinumerot ovat jo intuitiivisestikin ajateltuna toisistaan riippuvaisia, eivätkä ne yksinään muodosta järkevää osoitetietoa. Puhelinnumerot sen sijaan voivat esiintyä useammin yksinkin, mutta silti laskennallinen todennäköisyys on merkittävästi suurempi kuin todellinen. Mielestäni tämä johtuu siitä, että yhteystiedot halutaan yleensä julkaista mahdollisimman monipuolisesti.

Luvussa 2.2 mainitun yhdysvaltalaistutkimuksen mukaan noin 4,5% kaikista sivuista sisältää tunnistettavan yhdysvaltalaisen postinumeron, 8,5% sisältää tunnistettavan puhelinnumeron ja 9,5% sisältää ainakin toisen näistä [13]. Vertailun vuoksi tein vastaavan analyysin omalle tutkimusmateriaalilleni. Taulukosta 7 voidaan laskea, että mahdollisia postinumeroita on löytynyt 16,4 prosentista ja puhelinnumeroita 14,8 prosentista kaikista tutkituista dokumenteista. Ainakin toisen numeroista sisältää 20,5 prosenttia dokumenteista. Tutkimuksessa [13] tulokset edustavat tutkijoiden mielestä lukumäärien alarajaa johtuen käytetystä tutkimusmenetelmästä, joka hyväksyy vain niin sanotut varmat tapaukset. Tämän tutkimuksen kyseessä ollen ainakin postinumero edustaa varmasti lukumäärien ylärajaa, koska lukumäärään on hyväksytty kaikki viisinumeroiset luvut, jotka löytyvät postinumeroluettelosta. Postinumeroiden riippuvuus muista paikkatietoelementeistä viittaa kuitenkin siihen, että suurin osa postinumeroiksi tulkituista merkkijonoista todella on postinumeroita. Mikäli näin ei olisi postinumeroita esiintyisi paljon enemmän myös yksinään.

4.5.3 Dokumentin koko ja löytyneet paikkatietoelementit

Taulukossa 8 on esitelty löydettyjen paikkatietoelementtien lukumäärää tutkitun dokumentin kokoon suhteutettuna. Intuitiivinen perusoletus siitä, että kooltaan suurempi dokumentti sisältää enemmän paikkatietoelementtejä kuin pienempi dokumentti näyttää pitävän paikkaansa. Pienistä alle yhden kilotavun dokumenteista yli 90 % ei sisältänyt ainoatakaan paikkatietoelementtiä. Yli 10 kilotavun kokoisista dokumenteista jo kaksi kolmasosaa sisältää ainakin yhden paikkatietoelementin.

Taulukko 8. Löydettyjen paikkatietoelementtien lukumäärä dokumentin kokoon suhteutettuna.

	Löydetyt paikkatietoelementit (kpl)											
Koko	1	2	3	4	5	6-7	8-10	11-15	16-20	21-50	51-100	+100
0-500 t	88	27	27	5	3	2						
500-1000 t	321	125	96	45	36	7	5	2				
1-2 kt	327	217	196	119	72	62	16	9				
2-5 kt	502	379	351	329	243	220	124	60	21	5		
5-10 kt	405	393	422	317	305	239	198	88	25	23	3	1
10-20 kt	256	193	192	183	141	213	94	81	37	36	8	3
20-50 kt	78	50	61	49	48	61	45	49	24	30	4	5
50-100 kt	7	7	5	5	2	7	5	6	4	10	4	6
+100 kt	5	0	0	1	2	1	2	2	0	2		
Yhteensä	1989	1391	1350	1053	852	812	489	297	111	106	19	15

4.5.4 Harrastelijoiden ylläpitämien sivustojen sisältämät paikkatietoelementit

Tutkittaessa fi-päätteisiä web-sivustoja voidaan olettaa, että suurin osa on virallisten tahojen, kuten liikemaailman tai julkisten palveluiden verkkopalveluita. Tämä oletus voidaan tehdä, koska fi-päätteiset osoitteet olivat pitkään rajoitettuna vain tarkkaan määritellyille palveluille ja tuotemerkeille. Tutkin asiaa käytännössä käymällä läpi 3812 harrastelijoiden ylläpitämää sivustoa, jotka on poimittu *Mikrobitti*-lehden ylläpitämästä *koti.mbnet.fi*-palvelusta, johon lukijat voivat tuottaa www-sivuja. Tutkitut sivustot edustavatkin monenlaisia harrastuksia, henkilökohtaisia kotisivuja ja joukossa on muutamia pienyritystenkin sivuja.

Taulukossa 9 on eritelty epävirallisilta web-sivustoilta löydettyjä paikkatietoelementtejä. Vähintään yksi paikkatietoelementti löytyi 13,7 % sivustoista. Verrattaessa lukumääriä virallisilta sivustoilta löydettyihin vastaaviin lukumääriin voidaan todeta, että harrastelijoiden ylläpitämät epäviralliset web-sivustot sisältävät merkittävästi vähemmän paikkatietoelementtejä kuin viralliset sivustot.

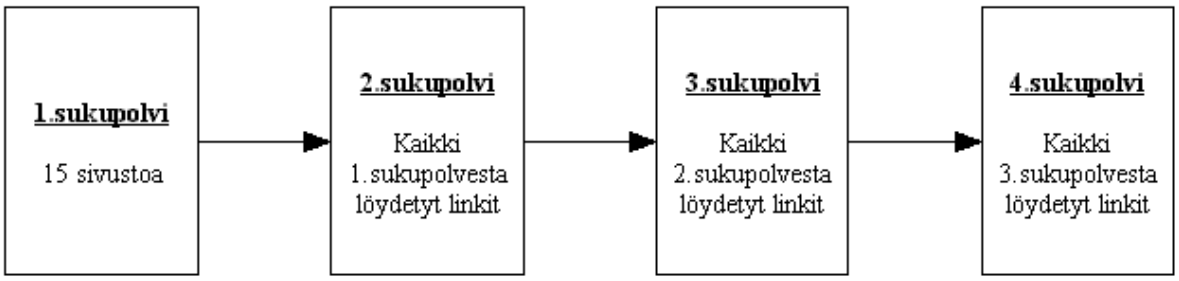
Taulukko 9. Harrastelijoiden ylläpitämien ja virallisten sivustojen sisältämät paikkatietoelementit.

	Harrastelijat			Viralliset
Elementti	Lukumäärä	Osuus		Osuus
Paikannimi	467	12,3 %		27 %
Kadunnimi	90	2,4 %		17 %
Postinumero	63	1,7 %		16 %
Puhelinnumero	107	2,8 %		15 %
Vähintään yksi	524	13,7 %		35 %

Tarkasteltaessa koko WWW:tä suomalaisten sivustojen osalta voidaan päätellä, että kerättävissä olevan paikkatiedon määrä mitä ilmeisimmin on harrastelijoiden ylläpitämiltä sivustoilta löydettyjen 13,7 %:n ja virallisempien tahojen 35%:n välillä. Luokittelu harrastelijoiden ylläpitämiin ja virallisiin sivustoihin on keinotekoinen, eikä sivustosta voi aina silmämääräisesti nähdä onko kyseessä harrastelijan ylläpitämä sivusto. Jo luokittelun vaikeuden takia on mahdotonta sanoa kuinka suuri osa verkon sisällöstä on harrastelijoiden tuottamaa ja kuinka suuri osa virallisempaa materiaalia. Mielestäni tulokset kuitenkin tukevat oletusta, että liikemaailman ja virallisten tahojen sivustot sisältävät enemmän hyödynnettävässä muodossa olevaa paikkatietoa kuin harrastelijoiden sivustot.

4.5.5 Leveys- ja syvyyssuuntainen verkon läpikäynti

Verkon läpikäyntisuunta voi vaikuttaa siihen, kuinka paljon paikkatietoa löydetään. Koska tässä tutkimuksessa on käyty läpi yli 24 000 sivustoa leveyssuunnassa, on tulosten arviointia varten tarpeen tehdä vertaileva läpikäynti myös syvyyssuunnassa. Valitsin tutkittaviksi sattumanvaraisesti 15 suomalaista verkkosivustoa, joiden osoitteen pääte oli fi, net tai com. Mukana oli sivustoja liikemaailmasta, harrastelijoilta ja valtion laitoksista. Läpikäynti suoritettiin kuvan 20 mukaisesti neljänteen sukupolveen asti. Yhteensä tutkittuja sivuja oli 4114.



Kuva 20. Syvyyssuuntaisen dokumenttien läpikäynnin neljä sukupolvea.

Syvyyssuuntaisessa läpikäynnissä tutkituista dokumenteista löydettyjen paikkatietoelementtien esiintymistiheydet olivat hyvin samansuuntaiset kuin leveyssuuntaisessa läpikäynnissä. Paikannimiä löytyi suhteellisesti enemmän kuin leveysläpikäynnissä, kun muiden elementtien esiintyminen oli hieman harvinaisempaa. Vaikka suhteellinen ero onkin varsin suuri, en pidä eroa merkittävänä, koska syvyyssuuntaisesta läpikäynnistä saatavat tulokset ovat sattumanvaraisempia kuin leveyssuuntaisen läpikäynnin tulokset. Taulukossa 10 tulokset on ryhmitelty elementtityypeittäin.

Taulukko 10. Syvyyssuuntaisessa läpikäynnissä löytyneet paikkatietoelementit.

Elementti	Syvyysläpikäynti			Leveys
	Lukumäärä	Osuus		Osuus
Paikannimi	1649	40 %		27 %
Kadunnimi	567	14 %		17 %
Postinumero	536	13 %		16 %
Puhelinnumero	670	16 %		15 %
Vähintään yksi	1816	44 %		35 %

Syvyyssuuntaisen tutkinnan tulokset riippuvatkin hyvin paljon läpikäynnin ensimmäisen sukupolven dokumenttien valinnasta, eikä tulos edusta yhtä kattavasti verkon sisältöä kuin aiemmin esitelty leveyssuuntainen läpikäynti. Syvyyssuuntaisessa läpikäynnissä ei voi etukäteen määritellä tarkasti mihin osaan verkosta läpikäynti lopulta johtaa vaan harhaan joutumisen riski on olemassa. Tässä tapauksessa pysyttiin tuloksista päätelleen varsin hyvin suomalaisilla sivustoilla, johtuen osittain siitä että läpikäynti ei mennyt neljättä sukupolvea syvemmälle.

Leveys- ja syvyyssuuntainen läpikäynti näyttävät antavan varsin samanlaisia paikkatietoelementtien esiintymistiheyksiä. Tässä tutkielmassa leveyssuuntainen läpikäynti oli käytännöllisempi, koska etukäteen pystyi määrittelemään mitä sivustoja verkosta tutkitaan ja näin varmistui jo etukäteen, että lopputulos edustaa kattavasti fi-päätteisiä osoitteita. Mikäli tarkoituksena on vain kerätä paikkatietoa verkosta ei läpikäynnin suunnalla näyttäisi olevan suurta merkitystä.

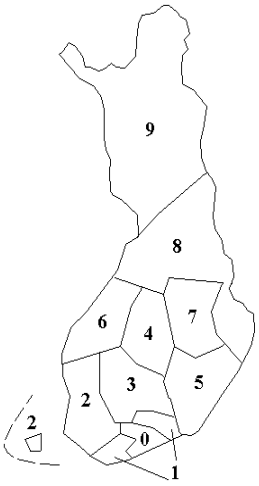
4.5.6 Paikkatietoelementtien painottuminen maantieteellisesti

Postinumeroiden ja paikannimien esiintymistiheyksiä tutkittaessa tein analyysin siitä, missä osissa Suomea verkkopalveluita on eniten ja millainen palveluiden jakauma on. Taulukossa 11 on tunnistettujen postinumeroiden jakautuminen postinumeroalueittain. Kuten oletettavaa onkin verkkopalvelut tiivistyvät postinumeroiden mukaan samoille alueille kuin asutuskin. Selvästi eniten tunnistettuja postinumeroita on pääkaupunkiseudulta. Muut keskittymät ovat Varsinais-Suomessa Turun alueella sekä Hämeessä.

Löytyneistä postinumeroista 42 % kohdistuu pääkaupunkiseudun 0-alueelle, jossa asuu 23 % Suomen väestöstä. Mielestäni tämä viittaa siihen, että pääkaupunkiseudun yritykset ovat

kiinnostuneempia toimimaan internetissä ja käyttävät internetiä yhtenä tiedotusvälineenä muiden joukossa. Suhteellisesti vähiten löytyneitä postinumeroita on Itä- ja Pohjois-Suomesta.

Taulukko 11. Löytyneiden postinumeroiden ja Suomen väestön jakautuminen alueittain [26].



Alue	Löydetyt postinumerot		Asukasluku	
0	2084	42 %	1216308	23 %
1	304	6 %	213561	4 %
2	620	13 %	713568	14 %
3	516	11 %	822399	16 %
4	255	5 %	266082	5 %
5	233	5 %	484259	9 %
6	325	7 %	437649	8 %
7	168	3 %	251356	5 %
8	148	3 %	428493	8 %
9	253	5 %	386057	7 %
Yht	4906	100 %	5219732	100 %

Vastaavanlaiset keskittymät löytyvät myös tunnistetuista paikannimista. Taulukossa 12 ovat kymmenen useimmin tunnistettua paikannimeä esiintymislukumäärineen. Selvästi eniten on löytynyt *Helsinki* eri taivutusmuodoissaan. Toiseksi yleisin paikannimi on *Tampere* ja kolmanneksi *Oulu*.

Taulukko 12. Yleisimmät paikannimet tutkituissa nettidokumenteissa

Paikannimi	**Lkm**
Helsinki	2477
Helsingin	1311
Tampereen	693
Tampere	647
Oulun	540
Turku	529
Espoo	386
Vantaa	350
Turun	350
Oulu	342

5 Yhteenveto

Web-dokumentti sisältää usein tietoa, jolla on maantieteellinen luonne. Tällainen tieto voi olla kirjoittajan tietoisesti dokumenttiin lisäämää tarkkaa sijaintitietoa, jolloin puhutaan eksaktista sijaintitiedosta. Yleisempi tilanne on, että sijaintitieto joudutaan johtamaan jostain muusta tiedosta. Tällaisia tapauksia ovat esimerkiksi puhelinnumerot ja osoitteet, joiden koordinaattitasoiseen kohdistamiseen tarvitaan paikkatietorekisteri. Rekisteri sisältää osoitteen ja vastaavan koordinaattiparin, jolloin tarkka kohdistaminen voidaan suorittaa. Puhelinnumeron kohdistaminen suoritetaan liittymän omistajan osoitetiedon perusteella. Eksaktin ja epäsuoran sijaintitiedon lisäksi kolmas paikkatiedon esiintymismuoto ovat pääteltävät elementit. Niitä ovat muun muassa valuutat, jotka voidaan vaihtelevalla todennäköisyydellä kohdistaa suuremmalle maantieteelliselle alueelle.

Internetistä löytyy useita paikkatietorekistereitä käyttäviä sovelluksia. Yksi tunnetuimmista lienee Keltaisten Sivujen verkkopalvelu, jonka reitti- ja karttahakujen ytimenä toimii osoiterekisteri. Google Local on paikallishakukone, jonka avulla voi etsiä pohjoisamerikkalaisia palveluita maantieteellisen sijainnin perusteella. Google Local on sekin rakennettu paikkatietorekisterin varaan, eikä avoimia koko verkon sisältöön hakuja tekeviä paikallishakukoneita ole käytössä.

Olen tutkinut tässä tutkielmassa paikkatiedon esiintymistiheyksiä suomalaisilla web-sivustoilla. Otokseen on valikoitunut runsaat 24000 fi-päätteisellä osoitteella varustettua sivustoa noin 86000:sta mahdollisesta. Jokaisesta sivustosta on tutkittu ainoastaan etusivu. Tästä johtuen tutkimustulokset eivät edusta kattavasti kaikkia suomalaisia verkkopalveluita maailmanlaajuisuudesta puhumattakaan, mutta tarjoavat kuitenkin suuntaa-antavia tuloksia siitä, kuinka paljon paikkatietoa suomalaiset verkkosivustot sisältävät.

Verkkosivuilta on tutkittu paikannimien, kadunnimien, postinumeroiden, puhelinnumeroiden, osoite-elementtien ja geo-tagien esiintymistä. Käytännön läpikäyntiä varten olen toteuttanut sovelluksen, joka suorittaa tiedon keräämisen WWW-sivuilta. Sivuilta etsitään ennalta määrättyjä 300:a paikannimeä, jotka edustavat Suomen suurimpia asutuskeskuksia. Kadunnimien, postinumeroiden ja puhelinnumeroiden etsiminen suoritetaan merkkijonojen perusteella. Sovelluksen lievästä suppeudesta johtuen osa olemassa olevasta paikkatiedosta jää löytymättä ja toisaalta sovellus löytää elementtejä, jotka eivät sisällä paikkatietoa, vaikka muodoltaan näyttävätkin siltä.

Esiintymistiheydeltään yleisimpiä paikkatietoelementtejä olivat paikannimet, joita löytyi noin 21000. Puhelinnumeroita löytyi runsaat 8000, kadunnimiä vajaat 6000 ja postinumeroita noin 5000. Paikkatietoa sisältäviä addres-tagilla varustettuja osoite-elementtejä tutkimuksessa löytyi vain kuusi kappaletta ja geo-tageja ei ainoatakaan. Web-dokumenttien osoitteista löytyi lisäksi 1500 paikkatietoa sisältävää osaa.

Paikkatietoelementit ovat usein riippuvaisia toisistaan. Toisin sanoen niille on luonteenomaista esiintyä erilaisissa ryhmissä. Esimerkiksi kadunnimet ja postinumerot esiintyvät yleensä yhdessä. Olen todennut tällaisten riippuvuuksien olemassaolon käytännössä tutkimalla löytyneiden paikkatietoelementtien esiintymistodennäköisyyksiä.

Paikkatiedon löytymistä edesauttaa, jos tutkittava sivusto on liikemaailman tai julkisen sektorin ylläpitämä, koska tällöin yhteystietojen löytyminen on todennäköisempää kuin harrastelijoiden sivustoilta. Olen tutkinut noin 4000 harrastelijoiden ylläpitämää sivustoa, joista löytyneiden paikkatietoelementtien määrä on huomattavasti pienempi kuin vastaava osuus liikemaailman ja julkisten palveluiden verkkopalveluissa.

Verkon tutkimisessa käytettävällä leveys- tai syvyyssuuntaisella läpikäyntialgoritmilla ei näytä olevan merkittävää vaikutusta paikkatiedon löytymiseen. Tässä tutkielmassa on käytetty leveyssuuntaista läpikäyntiä, koska se antaa mahdollisuuden etukäteen määritellä läpikäytävät sivustot toisin kuin syvyyssuuntainen läpikäynti, joka ohjautuu sattumanvaraisemmin. Läpikäyntialgoritmin valintaan en osaa antaa yleispätevää ohjetta, vaan valintaa tehtäessä on arvioitava läpikäynnin tavoitteita ja algoritmien soveltuvuutta tavoitteiden saavuttamiseen.

Postinumeroille ja paikannimille tehdyn analyysin mukaan web-dokumenttien sisältämät osoitetiedot keskittyvät samoille alueille kuin asutuskin eli pääkaupunkiseudulle, Turun seudulle ja Tampereen seudulle. Helsinki eri taivutusmuodoissaan on selkeästi yleisimmin käytetty paikannimi suomalaisilla verkkosivustoilla. Vaikka pääkaupunkiseudulla asuukin suuri osa suomalaisista, näyttävät web-dokumentit painottuvan sinne vieläkin selvemmin. Jopa 42% löytyneistä postinumeroista kohdistuu pääkaupunkiseudun 0-alueelle.

Paikkatiedon esiintymistiheyksiä tutkimalla voidaan tehdä päätelmiä, riittääkö kerättävissä olevan materiaalin määrä pohjaksi mahdollisten lisäsovellusten toteuttamiseen. Tämän tutkimuksen mukaan suomalaisista fi-päätteisistä osoitteista kerätyistä dokumenteista 35 prosenttia sisältää jotain

tunnistettavissa olevaa paikkatietoa. 20 prosenttia dokumenteista sisältää joko postinumeron tai puhelinnumeron. Mikäli tutkimus olisi kohdistunut kaikkiin suomalaisiin web-dokumentteihin, olisivat vastaavat luvut olleet mitä ilmeisimmin hieman pienempiä, kuten harrastelijoiden sivustoille tehty läpikäynti osoittaa. Kuitenkin lukuja arvioitaessa täytyy muistaa, ettei kaikilla verkossa julkaistuilla dokumenteilla missään tapauksessa ole maantieteellistä luonnetta, eikä kaikista dokumenteista mitenkään voi löytää viittauksia johonkin maantieteelliseen alueeseen.

Perustavanlaatuisena ongelma arvioitaessa tutkimustulosten jatkosoveltamismahdollisuuksia on ettei tämä tutkimus ota kantaa siihen ovatko dokumentit relevantteja niillä maantieteellisillä alueilla, joihin niistä löydetyt paikkatietoelementit viittaavat. Esimerkiksi se, että dokumentista löytyy sana *Helsinki* ei välttämättä tarkoita sitä, että kyseinen dokumentti välttämättä sisältää tietoa Helsingistä tai helsinkiläisistä palveluista. Tästä tilanteesta johtuen ei voida olla täysin varmoja että dokumentin kohdistus löytyneiden paikkatietoelementtien avulla todella onnistuisi. Paikkatietoelementtien riippuvuudesta tehdyt päätelmät kuitenkin osoittavat, että löytyneet elementit melko luotettavasti sisältävät sitä informaatiota jota niiden on oletettukin sisältävän. Mikäli elementtien esiintymisissä ei olisi havaittu riippuvuuksia, niiden soveltaminen olisi kyseenalaista.

Eräs sovellus, johon tässä tutkielmassa esiteltyä paikkatiedon eristämistä voisi käyttää olisi paikallisen hakukoneen toteuttaminen. Luvussa 3.1.4 esitelty Google Local antaa viitteitä millaisesta sovelluksesta voisi olla kyse. Erona Google Local:iin olisi, että Yellowpages:in tietokantojen sijaan hakualustana käytettäisiin koko verkkoa. Tässä tutkimuksessa 35 prosenttia tutkituista dokumenteista sisälsi paikkatietoelementtejä. Kyseinen luku kaikkia mahdollisia suomalaisia nettidokumentteja arvioitaessa lienee 15-30 prosenttia. Lukumäärä on kuitenkin niin suuri, että paikallisen hakukoneen toteuttaminen olisi järkevää, koska hakuja voitaisiin suorittaa tarpeeksi suuresta määrästä dokumentteja.

Paikkatiedon eristäminen nettidokumenteista voi toimia pohjana myös muille sovelluksille. Internet sisältää valtaisan määrän tietoa, jolla on maantieteellinen luonne. Tätä tietomassaa tullaan tulevaisuudessa käyttämään sovelluksissa, joita itse ei välttämättä pysty edes kuvittelemaan. Mobiilitekniikan ja internetin sulautuessa entistäkin tiukemmin toisiinsa saatavilla olevan paikkatiedon käyttäminen tulee entistäkin kiinnostavammaksi.

VIITTELUETTELO

[1] Bikel D.M., Schwartz R., Weischedel R.M. 1999. An algorithm that Learns What's in a Name. *Machine Learning (Special Issue on NLP)*, **34**(1-3): 211-231.
(Saatavana myös: http://www.cis.upenn.edu/~dbikel/papers/algthatlearns.doc.pdf, 18.10.2004)

[2] Bilhaut F., Charnois T., Enjalbert P., Mathet Y. 2003. Geographic reference analysis for geographic document querying. *Analysis of Geographic References. Workshop Held at the HLT/NAACL Conference 2003*. (Toim. Kornai A., Sundheim B.), Association for Computational Linguistics, Edmonton, Alberta, Canada, 55-62.
(Saatavana myös: Http://www.metacarta.com/kornai/NAACL/WS9/Conf/ws904.pdf, 18.10.2004)

[3] Daviel K. 2003. *Geographic registration of HTML documents*.
Http://geotags.com/geo/draft-daviel-html-geo-tag-06.html (18.10.2004).

[4] Gravano L., Hatzivassiloglou V., Lichtenstein R. 2003. Categorizing Web Queries According to Geographical Locality. *Proceedings of the twelfth international conference on information and knowledge management*, ACM Press, New Orleans, LA, USA, 325-333.
(Saatavana myös: http://www1.cs.columbia.edu/~gravano/Papers/2003/cikm03.pdf, 18.10.2004)

[5] Google Inc. 2004. *Google*.
Http://www.google.com/ (18.10.2004).

[6] Google Inc. 2004. *Google Local Beta*.
Http://local.google.com (18.10.2004).

[7] International Organization of Standardization. 1998. *ISO 3166-2*.
Http://www.iso.ch/iso/en/prods-services/iso3166ma/04background-on-iso-3166/iso3166-2.html (18.5.2004).

[8] ITU-T. 2004. *International Telecommunication Union – Telecommunication Standardization Sector*.
Http://www.itu.int/ITU-T/ (18.10.2004).

[9] Korkea-aho M. 2001. *Location Information in the Internet*. Licentiate's thesis, Helsinki University of Technology, Helsinki.

[10] Korpela J., Linjama T. 2004. *XHMTL-Käsikirja*. Docendo Finland Oy, Porvoo.

[11] Leidner J.L., Sinclair G., Webber B. 2003. Grounding Spatial Named Entities for Information Extraction and Question Answering. *Analysis of Geographic References. Workshop Held at the HLT/NAACL Conference 2003*. (Toim. Kornai A., Sundheim B.), Association for Computational Linguistics, Edmonton, Alberta, Canada, 31-38.
(Saatavana myös: Http://www.metacarta.com/kornai/NAACL/WS9/Conf/ws908.pdf, 18.10.2004)

[12] Markowetz A., Brinkhoff T., Seeger B. 2004. *Geographic Information Retrieval. Workshop Held At the 3rd International Workshop on Web Dynamics at WWW2004*.
Http://www.dcs.bbk.ac.uk/webDyn3/webdyn3_proceedings.pdf (20.10.2004)

[13] McCurley K.S. 2001. Geospatial Mapping and Navigation of the Web. *Proceedings of the tenth international conference on World Wide Web*, ACM Press, New York, NY, USA, 221-229.
(Saatavana myös: Http://www10.org/cdrom/papers/278/, 18.10.2004)

[14] Mikheev A., Moens M., Grover C. 1999. Named Entity Recognition without Gazetteers. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Bergen, Norway, 1-8.
(Saatavana myös: Http://acl.ldc.upenn.edu/E/E99/E99-1001.pdf, 18.10.2004)

[15] Moorman K. 1999. Web review: W3C, the World Wide Web consortium. *Crossroads (Special issue on markup languages)*, **6**(2): 4.

[16] Morimoto Y., Anono M., Houle M., McCurley K. 2003. Extracting Spatial Knowledge from the Web. *Proceedings of the IEEE Symposium on Applications and the Internet*, IEEE, Orlando, Florida, USA, 326-333.
(Saatavana myös: Http://www.almaden.ibm.com/cs/people/mccurley/pdfs/SAINT03.pdf, 18.10.2004)

[17] National Geospatial Intelligence Agency. 2004. *GEOnet Names Server*.
Http://earth-info.nga.mil/gns/html/index.html (18.10.2004).

[18] National Geospatial Intelligence Agency. 2004. *National Geospatial Intelligence Agency*. Http://www.nga.mil/portal/site/nga01/ (18.10.2004).

[19] Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. 1999-2004 . *Alexandria Digital Library Gazetteer*. Http://www.alexandria.ucsb.edu/gazetteer (18.10.2004).

[20] Scheaffer R. 1990. *Introduction to Probability and its Applications*. PWS-Kent Publishing, Boston.

[21] Suomen Keltaiset Sivut Oy. 2004. *Suomen Keltaiset Sivut*. Http://www.keltaisetsivut.fi (18.10.2004).

[22] Suomen Posti Oyj. 2004. *Postinumeroluettelo*. Http://www.posti.fi/postinumeroluettelo/ (18.10.2004)

[23] Suomen Viestintävirasto. 2003. *Suomen viestintäviraston verkkopalvelu*. Http://www.ficora.fi/ (18.10.2004).

[24] Tekniikan Sanastokeskus. 2002. *Paikannussanasto.* Http://www.tsk.fi/fi/info/paikannussanasto.pdf (18.10.2003).

[25] Tilastokeskus. 2003. *Tietotekniikka kotitalouksissa 1990-2001*. Http://statfin.stat.fi/StatWeb/start.asp?PA=Erlaityl&D1=a&D2=a&LA=fi&DM=SLFI&TT=2 (18.10.2004).

[26] Tilastokeskus. 2003. *Väestö iän mukaan alueittain 1980-2003*. Http://statfin.stat.fi/statweb/catnewfi/Vik9802.asp (23.11.2004).

[27] Treese W. 1999. Putting it together: Engineering the net: the IETF. *NetWorker*, **3**(1):13-19.

[28] Zhou G. Su J. 2002. Named Entity Recognition using an HMM-based Chunk Tagger. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, Philadelphia, USA, 473-480.
(Saatavana myös: http://acl.ldc.upenn.edu/P/P02/P02-1060.pdf, 18.10.2004)

[29] Zhou G., Su J. 2003. Integrating Various Features in Hidden Markov Model Using Constraint Relaxation Algorithm for Recognition of Named Entities without Gazetteers. *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering,* IEEE, Beijing, China, 465-470.