Kimmo Tuominen

Tarvitaanko verkkodokumenttien kuvailussa käsityötä?

Osa 1: kontekstuaalinen metadata

Kimmo Tuominen, Tarvitaanko verkkodokumenttien kuvailussa käsityötä? Osa 1: kontekstuaalinen metadata. [Do we need human-made descriptions of Web documents? Part 1: contextual metadata] Informaatiotutkimus 22(3), 87-94.

The first part of the paper deals with the issues of document authenticity, cognitive authority and multiperspectivity in the Web. By using contextual metadata in search results a search engine or a digital library can give users the kind of hints or clues about the nature of the described document or Web site that could not be generated by using ordinary content indexing technologies. These hints are about document's context: its origin, history, cognitive authority and place in an intellectual tradition. The author argues that this kind of contextual metadata should not be generated by machines, but by humans. Problems associated with the authenticity of contextual metadata are briefly discussed.

Address: Kimmo Tuominen, Department of Communication, FIN-00014 University of Helsinki, Finland. Email: Kimmo.Tuominen@helsinki.fi

Johdanto

Lähivuosina digitaalisen materiaalin sisältöön perustuvat laskennalliset tekniikat tulevat olemaan kaikkialla läsnä olevia, ensisijaisia indeksoinnin ja tiedonhaun välineitä: halpoja ja nopeita (Lynch 2000a). Jo nyt uusi verkkodokumentti voi tulla automaattisesti indeksoiduksi heti, kun se julkaistaan ensimmäistä kertaa. Vaikka luonnollisella kielellä operoivat hakukäyttöliittymät ovat vielä suhteellisen alkeellisia, monet toiminnot alkavat olla jo mahdollisia: Sisältölähtöisten tekniikoiden avulla voidaan erotella ihmisten, paikkojen ja organisaatioiden nimiä tekstimassasta sekä päätellä vakiintuneimpia dokumenttigenrejä. Hakukoneet ja uusien hakuvälineiden prototyyppiversiot muun muassa oikovat hakusanojen kirjoitusvirheitä sekä tarjoavat toisenlaisia kirjoitusvaihtoehtoja, laajentavat hakua synonyymisanastojen ja sanakirjojen avulla, antavat mahdollisuuden täsmentää hakulauseketta monitulkintaisen termin tarkempia merkityksiä määrittämällä sekä analysoivat muiden käyttäjien hakuja tarjotakseen rinnakkaisia hakulauseita tai lupaavalta vaikuttavia uusia termejä. Kehittyneimmät relevanssilajittelun algoritmit laskevat, termiesiintymien määrään ja sijaintiin perustuvan automaattisen analyysin lisäksi, verkkodokumentteihin tehtyjen ulkopuolisten linkkien määriä tai dokumenttien suosiota aiemmissa tiedonhauissa. Hakutuloksen esittämiseen keskittyvät ohjelmakomponentit puolestaan hyödyntävät erilaisia dokumenttiluokituksia tai lennossa tehtyjä klusterointeja sekä visualisoivat käsite- ja dokumenttiavaruuksia.

Indeksoinnin, tiedonhaun sekä hakutulosten jäsentämisen ja esittämisen teknologioiden kehityksen nykyvaiheessa tulee pohtia entistä tarkemmin ihmisten ja koneiden roolia metadatan[1] tuottamisessa. On olemassa tiedonhaun menetelmiä (esimerkiksi fraasihaut laajoista kokoteksteistä), jotka ovat mahdollisia ainoastaan automaattisten indeksointimenetelmien ansios-

ta. Vastaavasti voidaan kysyä, liittyykö digitaalisten dokumenttien kuvailuun ja analysointiin sellaisia osatehtäviä, joista ihminen suoriutuu konetta paremmin. Entä ovatko nämä tehtävät niin tärkeitä, että kalliin ja suhteellisen hitaan ihmistyövoiman käyttö on perusteltua sekä oikeutettua? Clifford Lynch (2000a; 2000b; 2001; 2002) on useissa kirjoituksissaan esittänyt tämän kaltaisia kysymyksiä. Hänen mukaansa automaattinen indeksointi ja käsityö[2] tulee nähdä toisiaan täydentävinä pikemminkin kuin kilpailevina strategioina: niin älyllisellä kuin koneellisellakin analyysillä on omat vahvuutensa.

Laatimassani kaksiosaisessa artikkelikokonaisuudessa tarkastelun kohteena on käsityön tarve verkkodokumenttien kuvailussa[3]. Kutsun dokumentin kontekstin – sen alkuperän, historian, tiedollisen auktoriteetin ja sitä ympäröivän keskustelullisen maaston – kuvailuja *kontekstuaaliseksi metadataksi*. Kysyn, edellyttääkö nimenomaan kontekstuaalinen metadata ihmisälyn käyttöä verkkodokumenttien analyysiin koneälyn sijasta. Jos – kuten pyrin todistamaan – vastaus on kyllä, onko kontekstuaalinen metadata niin tärkeä asia, että ihmistyövoiman käyttö on perusteltavissa? Kolmas tutkimuskysymykseni on se, millaisia mahdollisuuksia ja ongelmia ihmisten tuottaman kontekstuaalisen metadatan hyödyntämiseen liittyy.

Tässä artikkelikokonaisuuden alkuosassa koetan vastata yllä luetelluista kysymyksistä kahteen ensimmäiseen. Verkkoympäristö asettaa omanlaisiaan haasteita dokumenttien sekä metadatan autenttisuuden ja luotettavuuden selvittelylle, ja näiden haasteiden tunnistaminen on edellytys käsityön tarpeellisuuden teeman edelleen kehittelylle. Niinpä luonnehdin seuraavassa autenttisuuden, tiedollisen auktoriteetin ja moniperspektiivisyyden käsitteitä sekä sovellan niitä avoimeen verkkoympäristöön. Tutkimuskirjallisuuden lisäksi käytän myös havaintoaineistoa: todellisia tiedonhakuesimerkkejä. Artikkelikokonaisuuden jälkimmäisessä osassa siirrän autenttisuuden, luotettavuuden ja moniperspektiivisyyden puntaroinnin yhteisölliseen kontekstiin, jolloin myös kysymys käsityöhön perustuvan metadatan hyödyntämiseen liittyvistä mahdollisuuksista ja ongelmista asettuu paremmin vastattavaksi.

Yksi tapa tehdä karkea rajanveto käsityön ja automaattisen indeksoinnin välillä on puhua syvästä tai näkymättömästä verkosta: jatkuvasti päivittyvistä tietokannoista sekä kaiken aikaa lisääntyvästä ei-tekstuaalisesta verkkomateriaalista (videoista, musiikkista, jne.). Huolimatta esimerkiksi hahmon- ja puheentunnistusmenetelmien kehityksestä ihmiskäsin tuotetut tekstipohjaiset kuvailutiedot ovat olennaisen tärkeitä tämäntyyppisen – merkkijonoindeksointiin perustuvien sovellusten ulottumattomissa olevan – materiaalin haettavuuden mahdollistamisessa. Seuraavassa ei ole mahdollisuutta perehtyä tarkemmin näkymättömän verkon tematiikkaan, vaan rajoitan tarkasteluni kokotekstidokumentteihin: materiaaliin, jonka indeksoinnissa hakukoneet ovat tällä hetkellä vahvimmillaan.

Verkkosivujen ja -dokumenttien kirjastomaisen (rakenteiseen formaattiin ja tesauruksiin pohjautuvan) luetteloinnin tarpeellisuutta perustellaan usein sillä, että karkeat sanahaut tuottavat tulokseksi tuhansia roskaviitteitä. Hakukoneet eivät nykyään pysty erottelemaan esimerkiksi tekijää ja kohdehenkilöä, ja relevantin tiedon paikallistaminen verkkosälän seasta on muutenkin työlästä. Niinpä kirjastomaiset menetelmät voisivat laajemmin käytettyinä tehostaa ja nopeuttaa tiedonhakua. Seuraavassa en puutu myöskään tähän hakukoneiden puutteellisuutta koskevaan keskusteluun. Totean vain, että automaattiset tiedonhaun tarkkuutta ja tyhjentävyyttä lisäävät menetelmät kehittyvät jatkuvasti; esimerkiksi kieliteknologisten työkalujen luotettavuus dokumenttien sisällön analysoinnissa kohentunee entisestään lähivuosina.

Verkkodokumenttien ja metadatan autenttisuus

Verkon kontrolloimattomuus ja yhtäaikainen demokraattisuus sekä anarkistisuus ovat tuskin enää uutisia. Tiedämme, että verkkojulkaisuohjelman puutteellinenkin hallinta ja ilmaiseksi tarjolla oleva palvelintila mahdollistavat sen, että www:ssä kuka hyvänsä voi esimerkiksi julkaista ”uudelleenarviointeja” juutalaisten joukkomurhasta. Kyberavaruudessa ei ole mainoskatkoja, vaan faktojen esittämisen, propagandan, kaupallisen informaation ja poliittisen satiirin genrerajat ovat usein liukuvampia kuin painettujen dokumenttien maailmassa. Kun viruksetkin naamioituvat rakkauskirjeiksi työkaverilta, olemme tilanteessa, jossa digitaaliset identiteetit sekä näiden identiteettien tuottamien dokumenttien alkuperät ja tarkoitukset uhkaavat jäädä hämäriksi. Niinpä syste-

maattisen vilpin mahdollisuuden minimointi muodostuu verkkoympäristössä tärkeäksi, joskin vaikeasti saavutettavaksi, päämääräksi (Lynch 2001)[4].

Kaiken ihmiskäsin tuotetun metadatan hyödynnettävyyttä haittaa se, että avoimessa verkossa meillä ei ole takeita metadatan julkaisijan vilpittömyydestä (Lynch 2000b). Metadatan laajamittaisempi hyväksikäyttö on mahdollista suljetuissa ja kontrolloiduissa ympäristöissä, kuten perinteisissä viite- ja tekstitietokannoissa, joissa esiintyvien luettelointitietueiden laatijoiden identiteetti on melko yksiselitteisesti varmistettavissa ja heidän toimintansa on suhteellisesti ottaen luotettavaa. Sen sijaan avointa www:tä indeksoivien sovellusten kehittäjien on otettava huomioon se, että verkkodokumenttien tekijöillä voi olla vahvoja intressejä vaikuttaa siihen, miten dokumentti tulee indeksoiduksi ja kuinka ylhäällä se esiintyy hakutuloksissa. Metadatalla voidaan siis pyrkiä manipuloimaan hakukoneiden toimintaa (Lynch 2001).

Koska verkkojulkaisujen ja -metadatan hyödyntäminen on ongelmallista, viime aikoina on tehty paljon kehitystyötä erilaisten autenttisuuden, luotettavuuden ja eheyden todentamisen tekniikoiden, kuten tarkistussumma-algoritmien sekä digitaalisten allekirjoitusten, parissa (Lynch 2000b). Vielä ei kuitenkaan ole käytettävissä luotettavia automaattisia keinoja sen päättelemiseen, voidaanko esimerkiksi avoimessa verkossa olevaan metadataan luottaa vai onko se liitetty tiettyyn verkkosivuun, jotta indeksointiprosessia voidaan manipuloida (Lynch 2001). Kontrolloimattoman metadatan käytön riskit ovat niin suuria, että hakukoneet jättävät usein esimerkiksi HTML-dokumenttien meta-kentät suosiolla indeksoimatta.

Autenttisuus viittaa dokumentin, metadatan tai yksilön identiteetin alkuperään: kun henkilö, asia tai esine X on autenttinen, X on se, joka se näyttää tai väittääkin olevansa tai joka sen väitetään olevan. Tietystä dokumentista tehdyn kopion autenttisuus riippuu etenkin käyttökontekstista ja sosiaalisista konventioista: Wittgensteinin lähettämän kirjeen digitoitu versio voi olla riittävän autenttinen filosofian historiasta kiinnostuneelle maallikolle, mutta tämä versio ei kenties riitä salapoliisintyötä tekevälle elämäkerran kirjoittajalle, joka haluaa pidellä kirjettä käsissään ja tutkia sitä fyysisenä esineenä erilaisten johtolankojen löytämisen toivossa. Kopion ja originaalin suhde ei siis aina ole yksiselitteinen: yhteisölliset ja käytännölliset sopimukset määrittävät sen, mitä ominaisuuksia kopion on säilytettävä, jotta se tulisi luokitelluksi autenttiseksi (vrt. Gladney & Bennett 2003)[5]. On myös muistettava, että vaikka dokumentti olisi täydellisen autenttinen ja ehyt, se voi silti sisältää asiavirheitä, valheita ja typeryyksiä (Lynch 2000b). Dokumentin autenttisuus ei todista, että se tai sen laatija olisi luotettava tiedollinen auktoriteetti tai että sen sisältämät väitteet pitäisivät paikkaansa [6].

Tiedollinen auktoriteetti

Institute for Historical Reviewin (IHR:n) kotisivua[7] voitaneen pitää edellä luonnostellussa mielessä autenttisena verkkodokumenttina. Sangen todennäköisesti IHR on sekä fyysisessä että virtuaalisessa maailmassa toimiva rekisteröity organisaatio, jolla on nimetty johtaja, postiosoite ja jonka julkaisemalla lehdellä on avustajakuntaa ja kansainvälinen toimitusneuvosto. Kotisivun tarkastelijalle ei herää epäilystä siitä, etteikö sivusto olisi IHR:n tai sen alihankkijan rakentama ja etteivätkö siinä esiintyvät näkemykset noudattaisi IHR:n poliittis-ideologista linjaa. Google-hakukoneen[8] tapa muodostaa viitetieto IHR:n pääsivusta on seuraavanlainen:

Institute for **Historical Review**

Institute for **Historical Review**. **...** The **Institute** for **Historical Review** is non-ideological, non-political, and non-sectarian. It is **...**

Description: Site of the world's leading Holocaust denial organisation. Many articles from its journal (founded...

Category: Society > Issues > ... > Holocaust Denial
www.ihr.org/ - 17k - Cached - Similar pages

Hakutuloksen otsikon jälkeen Google esittää indeksoidulta sivulta automaattisesti poimitun virkkeen, jonka mukaan IHR on epäideologinen, poliittisesti kantaa ottamaton ja lahkolaisuutta kaihtava organisaatio. Kuvauskentässä (Description) esitetty luonnehdinta on täysin toisenlainen: IHR:n päätavoite on kieltää toisen maailmansodan edellä ja sen aikana tapahtunut juutalaisten joukkomurha. Kategoriakenttä (Category) puolestaan paljastaa, että kuvattu sivu sijoittuu yhteiskunnallisten puheenaiheitten alaluokkaan *Holocaust_Denial*.

IHR sanoutuu faktalehtisessään[9] irti joukkomurhan kieltäjän leimasta, eivätkä kuvaus- ja kategoriakenttien tiedot selvästikään ole peräisin kyseiseltä organisaatiolta. Niinpä viitteen seitsemällä rivillä elää kaksi vastakkaista versiota IHR:n luonteesta ja statuksesta. On selvää, että ainoastaan IHR:n neutraaliutta ja puolueettomuutta korostava versio on generoitu automaattisesti.

Kategoriakentässä esiintyvä luokitushierarkia on linkki, jota seuraamalla pääsee Google-hakemiston (Google Directory) kohtaan *Society > Issues > Race-Ethnic-Religious_Relations > Holocaust_Denial*. Pian paljastuu, että Googlen esittämät kuvaus- ja luokitustiedot ovat peräisin Open Directory Projectin (ODP:n) useiden satojen hakukoneiden ja portaalien käyttöön antamasta tietokannasta[10]. ODP[11] mainostaa olevansa maailman suurin ihmiskäsin tuotettu aihehakemisto verkossa. Googlen koostama IHR:n kotisivun viitetieto, johon on yhdistelty automaattisella indeksoinnilla poimittuja sanoja sekä ODP:stä peräisin olevaa kontekstuaalista metadataa, ei kyseenalaista IHR:n sivuston autenttisuutta. Sen sijaan tuo seitsemän rivin kuvaus horjuttaa IHR:n *tiedollista auktoriteettia*: ”revisionistien” esittämät väitteet leimautuvat moraalisesti ja epistemologisesti kiistanalaisiksi.

Tiedollinen auktoriteetti (cognitive authority) on tiettyyn viralliseen tai epäviralliseen kategoriaan luettavissa oleva ihminen tai muu tiedonlähde, jolla ajatellaan olevan tai jonka ajatellaan sisältävän tietämystä jostakin asiasta (Wilson 1983). Tiedollisen auktoriteetin tarpeellisuus selittyy sillä, ettei meillä ole aikaa ja resursseja tarkistaa empiirisesti kuin hyvin pieni osa niistä lukuisista tietoväitteistä, joita ympärillämme risteilee. Kun laboratorioiden tai hiukkaskiihdytinten hankinta ja varustelu ovat äärimmäisen kallista puuhaa ja kun avaruussukkulaankaan ei ihan kuka tahansa pääse, joudumme tekemään muiden tuottamien representaatioiden varassa erotteluja toden, harhaluulon ja valheen välillä. Suurin osa kaikesta tiedostamme jää väistämättä uskonasiaksi ja auktoriteettikysymykseksi.

Tiedolliset auktoriteetit ovat usein jonkin asiantuntijakategorian edustajia, henkilöitä, joiden institutionaalinen asema ja henkilökohtainen kokemus viittaavat poikkeuksellisiin epistemologisiin kykyihin tietyllä alueella (Potter 1996). Ajattelemme stereotyyppisesti, että lääkärit tietävät sairauksista ja niiden parantamisesta ja että pankinjohtajat osaavat neuvoa asuntolaina-asioissa. Toisaalta tiedollinen auktoriteetti ei ole vain ihmisyksilöihin liitetty ominaisuus, vaan esimerkiksi tietty kirja, lehtiartikkeli, instrumentti, tai organisaatio voidaan mieltää vakuuttavaksi tiedon takeeksi (Wilson 1983). Esimerkiksi yliopistoon ja tiedeinstituutioon assosioidaan usein kollektiivisluontoista tiedollista auktoriteettia, joka on suhteellisen riippumatonta yksittäisten tutkijoiden ominaisuuksista tai teoista[12].

Tiedollisen auktoriteetin saavuttaminen edellyttää aina *luottamusta* (Burbules 2001). Asiantuntijuus on henkilökohtaisen tiedollisen auktoriteetin saamiselle tai ansaitsemiselle välttämätön, vaan ei riittävä, ehto. Voimme hyvinkin olla sitä mieltä, että henkilö X on astrologian ammattilainen, mutta silti suhtautua epäillen niihin ennustuksiin, joita hän meille tähtikarttaa tulkitsemalla tekee (Wilson 1983). Tiettyä tekstiä lukiessa punnitsemme paitsi argumentaation sisällöllisiä ja muodollisia piirteitä, myös ja etenkin kirjoittajatahon uskottavuutta: kysymme, minkälaisia tietäjiä dokumentin laatijat ovat. Luottamus ei ole puhtaasti epistemologinen kysymys, vaan teemme laajempiakin päätelmiä kirjoittajien arvoista, etiikasta ja tietynlaisesta hyveellisyydestä. Mitä lähempänä dokumentin laatijan maailmankuva on omaamme, sitä helpompaa meidän on hyväksyä hänen tiedollinen auktoriteettinsa.

Henkilökohtaisen tiedollisen auktoriteetin ja tietyn dokumentin tai tietoväitteen uskottavuuden pönkittämistä ja purkamista voidaan tarkastella myös retoriselta kannalta (Potter 1996). Ei liene sattuma, että IHR kertoo pääsivullaan olevansa epäpoliittinen ja lahkolaisuutta kaihtava organisaatio: julistautumalla neutraaliksi IHR yrittää torjua siihen eri puolilta kohdistettuja syytöksiä[13]. Tällainen toiminta kuuluu uskottavuuden rakentamisen kulttuurisiin konventioihin: kyse on kielellisistä menetelmistä, joiden avulla pönkitetään tietynlaista versiota ja suojellaan sitä kritiikiltä joko etu- tai jälkikäteen. Lukijan luottamuksen herättämisen kannalta on usein hyvin edullista, jos dokumentti voi naamioida vaikuttamaan pyrkivän luonteensa neutraaliudeksi ja kykenee kuin ohimennen marssittamaan vaikutusvaltaisia tiedonlähteitä, kuten Nobel-palkittuja tiedemiehiä tai lääkärintakkiin sonnustautuneita henkilöitä, puhumaan puolestaan.

Moniperspektiivisyys

Organisatoristen ja muiden, mahdollisesti piilevien, yhteyksien selvittäminen on olennainen dokumentin ja sen laatijoiden tiedollisen auktori-

teetin puntaroinnin strategia verkkoympäristössä (Fritch & Cromwell 2001)[14]. Hyödyntämällä ODP:ltä peräisin olevaa kontekstuaalista metadataa Google antaa välineitä IHR:n ideologisten yhteyksien hahmottamiseen jo hakutuloksen esittämisvaiheessa. ODP-dataan perustuvaan Google-hakemiston *Holocaust_Denial* -luokkaan on muodostettu alakategoria *Opposing_Views*, joten kiistelevien tahojen näkemyserot käyvät hakemiston kautta asiaan tutustuville hyvin selviksi.

ODP:n kategoriassa *Holocaust_Denial > Opposing_Views* on 19 viitettä, kun taas *Holocaust_Denial* -kategoriassa on 11 viitettä. Samanlaiset vastustavien näkemyksen kategoriat löytyvät ODP:stä monille uskonnoille, kuten kristinuskolle, ateismille, buddhalaisuudelle, spiritismille ja skientologialle (*Society > Religion_and_Spirituality > Opposing_Views* sisältää yhteensä 1663 viitettä). Myös muun muassa homo- ja biseksuaalisuudelle, filosofiselle objektivismille, ympäristöliikkeelle ja vihreälle aatteelle, anarkismille, sosialismille, nationalismille sekä aborttikysymykselle löytyy ODP:stä vastustavien näkemysten kategoria. Nämä kategoriat ovat hyvin tarpeellisia, sillä yleensä kiihkeästi kiistelevien tahojen sivuilta ei löydy linkkejä vastustajan laatimiin verkkodokumentteihin[15]. ODP:n aineistonvalintapolitiikassa[16] moniperspektiivisyys, yhteiskunnallisen keskustelun kohteena olevan tapahtuman tai asiaintilan valaiseminen mahdollisemman monelta suunnalta, on tärkeä valintakriteeri[17] (vrt. Tuominen 2001, Tuominen, Talja & Savolainen 2003).

Vaikka käytössämme olisi automaattisia luokitusmenetelmiä sekä muita kehittyneitä indeksoinnin ja tiedon esittämisen työkaluja, moniperspektiivisyyden hahmottaminen edellyttäisi inhimillistä arviointia ja harkintaa. Tämä johtuu etenkin siitä, että keskustelua luonnehtivat jännitteet voivat olla implisiittisiä: tietoväite on usein retorisesti konstruoitu mitätöimään sen esittämistilanteessa julkilausumatonta vaihtoehtoisia tulkintaa. Koska argumentaatiokonteksti (Billig 1987) ei käy ilmi suoraan tekstistä, sisältöpohjaiset teknologiat eivät sitä myöskään kykene tavoittamaan.

Vitsi- ja huijaussivut

Digitaalisen dokumentin autenttisuus ja tiedollinen auktoriteetti ovat yhteydessä toisiinsa. Informaatiolukutaidon opettajien tavoin myös huijarit ovat perehtyneet verkkoympäristön tarjoamiin mahdollisuuksiin rakentaa tai purkaa dokumentin kognitiivista auktoriteettia sen todellista alkuperää hämärtämällä tai väärentämällä. Tämän kaltaiset alkuperäänsä salailevat valhetiedon tai epäinformaation lähteet voidaan luokitella 1) *vitsisivuiksi tai -dokumenteiksi* ja 2) *huijaussivuiksi tai -dokumenteiksi* (Piper 2000).

Presidentti George W. Bushin ja tämän neuvonantajien toimintaa irvailevat lukuisat verkkodokumentit, jotka vielä lisääntynevät vuoden 2004 vaalien alla, sopivat esimerkeiksi ensiksi mainitusta kategoriasta. Vitsisivut saattavat hämätä domainnimellään[18], mutta ne paljastavat todellisen luonteensa hyvin nopeasti. Tavoitteena on ilmeinen parodia, kevyt tai raskaampi herja, ja osa tämän kaltaisista sivuista on hyvin harmittomia. Toisaalta naljailun taustalla voi piillä myös vakavampia tarkoitusperiä: huumori ja ironia ovat tehokkaita retorisia aseita.

Huijaussivut muistuttavat raha- ja tauluvääre nöksiä: ne pyrkivät ”käymään oikeasta” ja peittämään todellisen alkuperänsä. Niiden laatijoiden päämääränä voi olla varastaa jonkin organisaation tai yksityishenkilön tiedollinen auktoriteetti. Näennäisautenttisilla verkkosivuilla voidaan esimerkiksi perustella sellaisia toimintakäytäntöjä tai -suosituksia, joita väärennyksen kohde todellisuudessa vastustaa. Linkityksiin, visuaalisuuteen ja tekstiin perustuvan www-retoriikan keinoin voidaan niin ikään rakentaa auktoriteetin vaikutelmaa olemattomalle organisaatiolle tai fiktiiviselle henkilölle. Myös sivuston laatijan todellisen identiteetin ja institutionaalisten kytkentöjen hämärtäminen tai kätkeminen on mahdollista. Huijaussivujen tai -dokumenttien taustalla voi olla taloudellisia pyyteitä sekä esimerkiksi poliittisia, eettisiä tai rasistisia motiiveja. Toisaalta tällaisia verkkodokumentteja kyhäillään myös silkasta pilailuntarpeesta ja petkuttamisen ilosta.

Todellisen huijauksen ja liioittelun rajaa voi olla vaikea määritellä: henkilö tai organisaatio voi eri keinoin esittää itsensä paljon huomattavampana tiedollisena auktoriteettina kuin mitä hän tai tämä onkaan. Valheellisten ylistävien arvostelulausuntojen esittäminen dokumentista X merkitsee katteetonta yritystä kasvattaa sen kognitiivista auktoriteettia. Oikein siteeratuilla mutta kontekstistaan irrotetuilla kehuilla on jo pitkään markkinoitu kirjoja ja elokuvia. Kyberavaruudessa myös kehujen väärentäminen on entistä helpompaa.

Tietomurto on tehokas autenttisuuden turmelemisen keino: www-palvelimeen luvatta

tunkeutuneet krakkerit voivat muuttaa autenttista verkkodokumenttia joko vitsi- tai huijausmielessä. Tällainen sivuston usein hetkellinen kaappaus voi tähdätä paitsi tiedollisen auktoriteetin varastamiseen jotakin tarkoitusta varten, myös olemassa olevan organisaation naurettavaksi tekemiseen ja sen uskottavuuden romahduttamiseen. Kutsumattomien vieraiden motiivina voi olla myös ylvästelynhalu: maineen kartuttaminen muiden krakkereiden viiteryhmässä.

Verkkodokumenttien autenttisuuden ja tiedollisen auktoriteetin tutkailu kysyy aikaa ja taitoa. Tietoresurssien arviointi ei useinkaan ole binaarinen prosessi, jossa käytettävissämme olisi kaksi arvoa, tosi ja ei-tosi, vaan joudumme operoimaan todennäköisyyksillä ja suhteellisuuksilla (Lynch 2000b). Viime kädessä kysymys on myös luottamuksesta sekä yhteisöllisistä uskomuksista ja päätöksistä (Lynch 2001). Esimerkiksi kryptokrafiset menetelmät (kuten digitaaliset allekirjoitukset) edellyttävät uskoa paitsi sertifikaattiteknologian toimivuuteen, myös siihen, että sertifikaatteja myöntävän tahon toiminta sekä yksityisen avaimen haltijan käyttäytyminen on luotettavaa (eli mikään toinen taho ei ole voinut saada vahingossa tai varastamalla yksityistä avainta haltuunsa). Teknisten tarkistusmekanismien käytön ohella myös ihmisälyyn turvautuminen on tarpeellista, kun halutaan ottaa kantaa näennäisautenttisuuden ja -auktoriteetin sekä moniperspektiivisyyden kysymyksiin.

Lopuksi

Tämän artikkelin alkupuolella kysyin, tarvitaanko kontekstuaalisen metadatan laatimisessa arviointia ja harkintaa vai pelkkää koneälyä. Toinen, edelliseen kytkeytyvä, tutkimuskysymykseni oli se, kuinka manuaalisesti tuotetun metadatan tarpeellisuutta voidaan perustella. Nämä kysymykset johdattivat autenttisuuden, tiedollisen auktoriteetin ja moniperspektiivisyyden problematiikan tarkasteluun avoimessa verkkoympäristössä. Ilmeni, että pelkkään kokoteksti-indeksointiin perustuvat ohjelmistot eivät voi ottaa tarpeeksi huomioon verkosta löytyvien dokumenttien keskustelullisia suhteita.

Edellä esimerkkinä käytetty Google sekä useat muut teknisesti kehittyneet hakukoneet ja portaalit hyödyntävät järjestelmissään myös ihmisaivojen ja -käsien tuottamia kuvailutietoja. Tämä selittyy osittain sillä, että manuaalisesti tuotetun metadatan avulla voidaan laajentaa tai tarkentaa hakua ennalta valittuun ja laadukkaaksi arvioituun aineistoon. Toinen – tämän artikkelin kannalta olennaisempi – perustelu käsityöhön turvautumiselle on kuitenkin se, että näin saadaan parhaiten esille myös dokumenttien intellektuaaliseen taustaan sekä dokumenttien autenttisuuteen ja tiedolliseen auktoriteettiin liittyviä, toisinaan tarkoituksella kätkettyjä, aspekteja. Näiden aspektien tai metakysymysten hahmottaminen puolestaan on ensiarvoista, kun tietoväitteitä arvioidaan ja perusteltuja tulkintoja luodaan. Tarvitsemme sekä ihmisvoimiin tuotettuun metadataan perustuvia että automaattista ja manuaalista luettelointia eri tavoin kombinoivia tiedonhakupalveluja.

Hyväksyttu julkaistavaksi 17.9.2003

Lähteet

Billig, M. (1987). *Arguing and thinking: a rhetorical approach to social psychology*. Cambridge: Cambridge University Press.

Burbules, N.C. (2001). Paradoxes of the Web: the ethical dimensions of credibility. *Library Trends* 49(3), 441-453.

Fritch, J.W. & Cromwell, R.L. (2001). Evaluating Internet resources: identity, affiliation, and cognitive authority in a networked world. *Journal of the American Society for Information Science and Technology* 52(6), 499-507.

Gladney, H.M. & Bennett, J.L. (2003). What do we mean by authentic? What's the Real McCoy? *DLib Magazine* 9(7/8). URL: http://www.dlib.org/dlib/july03/gladney/07gladney.html (30.7.2003).

Lynch, C.A. (2000a). The new context for bibliographic control in the new Millennium. Paper presented at the *Bicentennial Conference for the New Millennium: Confronting the Challenge of Networked resources and the Web*, Washington, DC, November 15-17, 2000. URL: http://lcweb.loc.gov/catdir/bibcontrol/lynch_paper.html (29.7.2003).

Lynch, C.A. (2000b). Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust. In *Authenticity in a digital environment*, 32-50. Washington, DC: Council on Library and Information Resources. URL: http://www.clir.org/pubs/reports/pub92/lynch.html (29.7.2003).

Lynch, C.A. (2001). When documents deceive: trust and provenance as new factors for information

retrieval in a tangled Web. *Journal of the American Society for Information Science and Technology* 52(1), 12-17.

Lynch, C.A. (2002). Digital collections, digital libraries, and the digitization of cultural heritage information. *First Monday* 7(5). URL: http://www.firstmonday.org/issues/issue7_5/lynch/index.html (29.7.2003).

Piper, P.S. (2000). Better read that again: Web hoaxes and misinformation. *Searcher* 8(8), 40-53. URL: http://www.infotoday.com/searcher/sep00/searcher.htm (17.8.2003).

Potter, J.(1996), *Representing reality: discourse, rhetoric and social construction*. London: Sage.

Preserving Digital Information (1996). Report of the Task Force on Archiving of Digital Information. Commissioned by The Commission on Preservation and Access and The Research Libraries Group, Inc., May 1, 1996. URL: http://www.rlg.org/ArchTF/tfadi.index.htm (28.7.2003).

Tuominen, K. (2001). *Tiedon muodostus ja virtuaalikirjaston rakentaminen: konstruktionistinen analyysi*. Espoo: CSC – Tieteellinen laskenta Oy. URL: http://acta.uta.fi/teos.phtml?5179 (12.8.2003).

Tuominen, K., Talja, S. & Savolainen, R. (2003). Multiperspective digital libraries: the implications of constructionism for the development of digital libraries. *Journal of the American Society for Information Science and Technology* 54(6), 561-569.

Wilson, P. (1983). *Second-hand knowledge: an inquiry into cognitive authority*. Westport: Greenwood Press.

Loppuviitteet

Kiitän Reijo Savolaista tämän artikkelikokonaisuuden viimeistelyä edistäneistä kommenteista.

[1] Metadata tarkoittaa verkkodokumentin tai muunlaisen informaatioresurssin rakenteista (yhteisesti sovittuun ja useimmiten myös standardoituun formaattiin nojautuvaa) kuvausta, jolla yritetään helpottaa tiedon hakua ja paikallistamista sekä dokumenttien relevanssin, käyttökelpoisuuden, luotettavuuden, laadun, uskottavuuden, alkuperän ja saatavuuden arviointia.

[2] Käytän ilmaisua ”käsityö” vastakohtana automaattiselle tekstisisällön indeksoinnille; kyseessä on metaforinen ilmaisu, joka viittaa ihmisten manuaalisesti tuottamaan metadataan. Toimintaprosessina käsityö on luonnollisesti sekä havaittavaa työskentelyä tietokoneen näppäimistöllä että aivotyötä: dokumentin sisällöllisten piirteiden ja sen kontekstin analysointia. ”Käsityö” on kuitenkin tässä yhteydessä osuva termi, sillä hyvä luetteloija ja sisällönkuvailija on myös hiljaisen tiedon taitaja samassa mielessä kuin vaikkapa kelloseppä. Toisaalta luetteloijakin voi tehdä huonoa käsityötä esimerkiksi poiketessaan perusteitta luettelointisäännöistä.

[3] Tiedon organisointia koskevien käsitteellisten tai paradigmaattisten innovaatioiden sekä kieliteknologian ja tiedon louhinnan menetelmien kehityksen myötä voimme joutua toistuvasti uudelleen määrittämään kone- ja ihmisluetteloijien rooleja. Niinpä analyysi koneen ja älyn välisestä rajanvedosta tiedonhakupalveluissa on aina historialliseen kontekstiinsa sidottu.

[4] Dokumenttien ja metadatan käsittelyssä voi syntyä myös virheisiin, tulkintaongelmiin ja tiedostojen korruptoitumiseen johtavia vahinkoja. Ne kuvastavat usein sekä verkkodokumenttien häilyvyyttä ja haavoittuvuutta että käyttäjien huolimattomuutta. Onnettomuuksien ja petosten seuraukset voivat olla aivan yhtä vakavia.

[5] Kun originaali A on valmiiksi digitaalinen dokumentti ja B on sen johdannaisdokumentti (jota on muovattu esimerkiksi pakkaus- tai tiivistämisalgoritmien avulla), kopion autenttisuutta ei voida tarkistaa eheyden (A:n ja B:n sisältämien bittisekvenssien identtisyyden) varmistavilla tarkistussummavertailuilla. B:n autenttisuuden määrittäminen on siis tällaisissakin tapauksissa konteksti- ja genre-sidonnainen asia. Esimerkiksi MP3-tiedostomuotoon tiivistetty musiikki tai MPEG1-tiedostomuodossa oleva elokuva saattaa riittää tavalliselle musiikin kuluttajalle tai videon katsojalle, mutta ei hifin harrastajalle tai DVD-tasoisen kuvan edellyttäjälle.

[6] Tiedot provenienssista voivat auttaa dokumentin tai kopion autenttisuuden varmistamisessa. Provenienssin dokumentoinnissa on kyse julkaisun alkuperän, olennaisten piirteiden ja elinkaaren kuvailusta. Provenienssitiedoista ilmenevät esimerkiksi dokumentin syntykonteksti, luontitarkoitus, omistus- ja valvontaketjut, tiedostomuunnokset, käyttöjärjestelmäkonversiot sekä versiohistoria. (Ks. Preserving… 1996.)

[7] http://www.ihr.org (14.8.2003).

[8] http://www.google.fi (5.8.2003).

[9] http://www.ihr.org/leaflets/fewfacts.html (5.8.2003).

[10] ODP-datan kierrätys selittää sen, miksi AlltheWeb-

ja Lycos-hakukoneet luonnehtivat IHR:n olemusta täsmälleen Googlen kuvauskentästä peräisin olevin sanakääntein, ks. http://www.alltheweb.com ja http://www.lycos.com (5.8.2003).

[11] http://dmoz.org (22.7.2003).

[12] Espoolaisessa Kauppakeskus Iso-Omenassa kuului elokuussa 2003 säännöllisen väliajoin luontaistuotekaupan mainos, jonka mukaan ”Turun yliopisto on todistanut tyrnimarjamehun terveellisyyden”.

[13] Ks. esimerkiksi osoitetta http://www.nizkor.org (14.8.2003).

[14]Tällainen toiminta ei ole tavatonta verkon ulkopuolellakaan, sillä kun tiedollista auktoriteettia puretaan, tulkintojen vääristyneisyyttä selitetään muun muassa havaintovirheillä, harhaluuloilla ja muilla intressitekijöillä (Potter 1996).

[15] IHR:n sivuilla on 14.8.2003 runsaasti linkkejä muihin ”revisionistisiin” organisaatioihin, mutta vain kolmeen ”antirevisionistiseksi” määriteltyyn tahoon (ja linkkien yläpuolella on maininta ”älä odota heidän linkittävän meidän sivuamme”). Termin ”antirevisionismi” käyttäminen on retorinen keino sävyttää linkkien tulkintaa: syntyvän vaikutelman mukaan ”revisionismi” olisi luonnollinen ajattelutapa ja ”antirevisionismi” puolestaan poikkeama normaalinäkemyksestä. IHR:n vastustajat eivät puhu revisionismista kuin korkeintaan lainausmerkeissä. Heidän retoriikassaan IHR määrittyy keskitysleirien olemassaolon kieltäjäksi. Yksittäiset sanat sekä niiden merkitykset ovat siis kiistelevien tahojen välisen kamppailun kohteita.

[16]http://dmoz.org/guidelines/include.html (17.8.2003).

17 Kaikki tässä ja edellisessä kappaleessa mainitut kategoriat ja viitemäärät on tarkistettu 17.8.2003.

[18]Muun muassa osoitteessa http://www.gwbush.com sijaitsee Yhdysvaltain istuvan presidentin herjaamiseen keskittynyt näennäisautenttinen verkkodokumentti (14.8.2003).