Kalervo Järvelin

Merkkijonot, sanat, termit ja käsitteet informaation haussa

Järvelin, Kalervo, Merkkijonot, Sanat, termit ja käsitteet informaation haussa [Strings, words, terms and concepts in information retrieval]. Kirjastotiede ja informatiikka 12 (4): 119–128, 1993.

In the literature and discussion about information retrieval (IR) the words "string", "word", "term" and "concept" are used inconsistently. There is no common agreement on the concepts named by these words. Thus the communication in IR is inefficient, inaccurate and prone to errors. Based on the concept of term in the study of terminology and on a three level model of IR, the paper seeks to clarify the terminology and proposes the terms string, search key and concept, with their subterms, for IR. Consequently it is proposed that a number of other words commonly used, such as "controlled term", "uncontrolled term", "free term", should not be used in IR.

Address: Department of Information Studies, University of Tampere, P.O.Box 607, FIN-33101, Tampere, Finland. Internet: kalervo.jarvelin@uta.fi

1. Johdanto

Informaation tallennuksen ja haun[1] kirjallisuudessa ja keskustelussa käytetään sanoja "merkkijono", "sana", "termi" ja "käsite" (engl. "string", "word", "term" ja "concept") huolimattomasti ja ilmeisen vaihtelevissa merkityksissä. Tällöin ei ole selvää, edustaako jonkin puhujan "sana" samaa käsitettä kuin jonkin toisen puhujan "sana" tai kolmannen "termi", "käsite" tai "merkkijono". Ei ole myöskään selvää, mikä tällöin on kunkin keskustelijan käsite sana. Seurauksena on, että informaation tallennuksen ja haun keskustelu on tehotonta ja epätarkkaa sekä johtaa helposti väärinymmärryksiin, joiden selvittely vaatii paljon selittämistä ja/tai tulkintaa.

Informaation tallennuksen ja haun keskustelu on tehotonta, koska keskustelijat l. puhujat ja kuulijat, kirjoittajat ja lukijat joutuvat näkemään vaivaa sopivan esityksen luomisessa ja tarkoitetun tulkinnan muodostamisessa. Se on epätarkkaa, koska keskustelijat eivät voi käyttää hyväksi erotteluja, joita merkkijonoihin, sanoihin, käsitteisiin ja termeihin voidaan liittää. Se johtaa väärinymmärryksiin, sillä kun jokainen tulkitsee määrittelemättömät ja huolimattomasti käytetyt sanat omalla tavallaan, ei mikään takaa tarkoitetun merkityksen välittymistä. Keskustelijan tekemä tulkinta tapahtuu usein vaistonvaraisesti hänen oman viitekehyksensä rajaamana. Silloin tällöin keskustelija joutuu kuitenkin luhistumistilanteeseen (Heideggerlaisittain break-down), joka havahduttaa miettimään oman tai toisen keskustelijan tarkoitusta.

Oppimisen edistämisessä, myös akateemisessa tutkimuksessa ja opetuksessa, on tärkeää pitää huolta käsitteistä ja kielestä. Ilman tällaista huolenpitoa on vaikea uskoa ymmärryksen kehittyvän informaation tallennuksen ja haunkaan alueella. Vain näin voidaan toivoa vältettävän tehottomuuden, epätarkkuuden ja väärinymmärtämisen vitsaukset. "Ohipuhumisesta" lienee jokseenkin jokaisella tarpeeksi kokemusta – terminologiaa pitää siis kehittää.

Seuraavassa pyrin aluksi tuomaan esille sanankäytön ongelman laajuuden informaation tallennuksen ja haun keskustelussa esimerkkien avulla.

Sitten tarkastelen sitä, miten ongelmaa voidaan jäsentää. Lopuksi teen ehdotuksen ongelman ratkaisuksi.

2. Sanankäytön ongelmia

Tämä jakso koostuu pääasiassa esimerkeistä, jotka ilmentävät sanojen "merkkijono", "sana", "termi" ja "käsite" käyttöä informaation tallennuksen ja haun keskustelussa. Osa esimerkeistä on aitoja – ne on siis poimittu minun tai muiden keskustelijoiden kirjoituksesta. Esimerkkien lähteitä en mainitse. Koska huolimattomia keskustelijoita ovat lähes kaikki, ei kunkin esimerkin nimenomaisella lähteellä ole suurtakaan merkitystä. Toisaalta esimerkit on irrotettu asiayhteydestään, joka sinänsä saattaa olla hyvin johdonmukainen ja huolellinen (mutta erilainen kuin muilla). Osa esimerkeistä on vain aidon näköisiä. Ne olen itse laatinut, mutta ne voisivat mielestäni esiintyä informaation tallennuksen ja haun keskustelussa. Lukija voikin miettiä, mitkä ilmaisut sopisivat myös omaan puheenparteen, mitkä taas aiheuttaisivat tulkinnan luhistumisen ja hämmästelyn. En erottele aitoja ja aidon näköisiä esimerkkejä toisistaan. Esimerkeissä sanoilla "merkkijono", "sana", "termi" ja "käsite" voi olla etuliitteitä, kuten "haku-", "indeksi-", "kuvailu-", "asia-" tai "vapaa-". Ne voivat esiintyä myös sanaliitoissa toisten sanojen, kuten "kontrolloitu" tai "kontrolloimaton", kanssa.

Ryhmä 1 : Termi vs. käsite

"Haun suunnittelu alkaa hakupyynnön analyysistä, jossa selvitetään pyynnön keskeiset termit, hakutermit."

"Hakupyynnön analyysin tarkoituksena on tunnistaa pyynnön keskeiset käsitteet."

Ryhmä 2 : Termi vs. sana vs. käsite

"Kun [dokumentin sisältämät] termit käännetään keinotekoiselle dokumentaatiokielelle, [...]."

Ryhmä 3 : Termi vs. sana

"Koska tekstissä on aina sanoja, jotka esiintyvät useasti, mutta jotka eivät ole merkityksellisiä, kaikki useasti esiintyvät termit eivät ole käyttökelpoisia automaattisessa indeksoinnissa. Tämäntyyppiset termit on kuitenkin mahdollista sulkea pois ns. STOP -listalla."

"SAP-projektissa [...] kirjaukseen otettiin mukaan dokumentin sisällysluettelo, valikoituja osia dokumentissa olevista taulukoista ja [...]. Näihin elementteihin voitiin kohdistaa vapaasanahaku."

"Suomenkielisiä hakutermejä käytettäessä on muistettava suomenkielen sanojen taipuminen. "

"[...] voidaan ajatella haun onnistumisen riippuvan myös siitä, miten hyvin ne termit, joita dokumenttien kuvailussa on käytetty, vastaavat niitä termejä, joita dokumenteissa tuodaan esiin tai miten hyvin ne vastaavat niitä termejä, joilla hakija aihetta lähestyy. "

"Luonnollista kieltä käytettäessä haun spesifisyys on riippuvainen sekä itse dokumentissa esiintyvien termien spesifisyydestä, esim. otsikossa ja abstrakteissa esiintyvien termien spesifisyydestä että haettaessa hakijan ajattelemista spesifeistä termeistä ja näiden yhteensopivuudesta dokumentissa käytettyjen termien kanssa. [...] Dokumentaatiokieltä käytettäessä spesifisyys on taas riippuvainen siitä, miten spesifejä termejä on valittu luokitettaessa ja indeksoitaessa. "

"Indeksoinnissa käytettävät sanat voivat olla avainsanoja tai asiasanoja. Asiasanat otetaan kontrolloidusta sanastosta (asiasanastosta, tesauruksesta), avainsanat ovat vapaasti valittavissa. "

"Asiasanaindeksoinnin tehtävä on siis tarjota sanoja, joilla kuvata dokumentin asiasisältöä. [...] Yksinkertaisemmassa asiasanastossa termit järjestetään aakkosiin ..."

"Tutustumisvaiheessa saatu idea dokumentin sisällöstä muutetaan sanalliseen muotoon, termeiksi, [...]. Tässä vaiheessa termit eivät vielä tarkoita dokumentaatiokielen termejä tai notaatioita, vaan yleensä termejä, [...]."

"[Dokumentaatiokielen rakentamisessa] yhtenä ryhmänä ovat termit, joilla on epätarkka merkitys, ja tämä epätarkkuus vaatii aina tulkintaa, esimerkkinä tämänlaatuisista termeistä voisi mainita termin ihmisoikeudet."

"Assosiaatiosuhteiden esittämisellä dokumentaatiokielessä halutaan hakijalle antaa vinkkejä siitä, millä muilla hakusanoilla hänen etsimäänsä aihetta voisi lähestyä."

"Asiakasta voi myös pyytää miettimään sopivia hakutermejä. [...] kysyy hakijan itsensä ehdottamia hakusanoja ja selvittää, minkätyyppiseen tarkoitukseen."

"Läheisyysoperaattoria tarvitaan varsinkin haettaessa tekstikentistä kontrolloimattomilla hakutermeillä."

Ryhmä 4 : Termi vs. sana vs. merkkijono

"Mikäli hakija haluaa etsiä viitteitä, joiden tietyssä kentässä hakutermi esiintyy, tulee hänen käyttää kenttätunnisteita."

"Hakuprofiilin laadinnassa on mietittävä myös hakutermien katkaisu."

"Jos hakusana katkaistaan liian lyhyeksi, se voi tuottaa hakutulokseen roskaa."

"Vektorimallissa tarkastellaan dokumenttien kuvailua termivektoreina. Kuvailutermit ovat dokumenttien tekstien sanoja, tiivistelmien sanoja tai dokumenteille annettuja dokumentaatiokielen termejä."

"Jos kaksi Boolen operaattoreilla yhdistettyä termiä löytyy pitkästä artikkelista, niillä ei välttämättä ole yhteyttä keskenään."

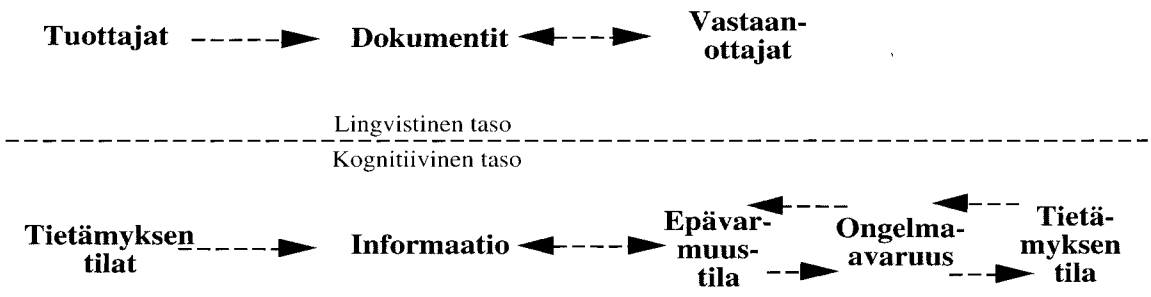
Asiantunteva lukija ymmärtää esimerkit oikein omissa asiayhteyksissään. Näyttää kuitenkin siltä, että sanat "merkkijono", "sana", "termi" ja "käsite" ovat informaation tallennuksen ja haun puheessa vain sanoja. Jokainen tarkoittaa niillä mitä haluaa. Niistä täytyy tehdä termejä, jotka ovat alalla vakiintuneita yleisesti hyväksyttyjä hyvin määriteltyjen käsitteiden nimityksiä. Käsitteet sana, termi ja käsite ovat informaation tallennuksen ja haun peruskäsitteitä, joilla voisi olettaa olevan selkeät määritelmät – mutta joko määritelmät on kirjoitettu?

3. Informaation tallennuksen ja haun tasot

Kommunikaatioprosessi

Kommunikaatiojärjestelmiä voidaan jäsentää informaatiotutkimusta varten kuvan 1 mukaisesti. Kuvan yläosassa, lingvistisellä tasolla, vasemmalla ovat tuottajat, jotka tuottavat dokumentteja (tekstejä), ja oikealla vastaanottajat, jotka hankkivat dokumentteja käyttöönsä. Alaosassa, kognitiivisella tasolla, vasemmalla esitetään tuottajien tietämyksen tila, josta kommunikoitavaksi tarkoitettu osa on dokumenttina esitettyä informaatiota. Oikealla esitetään vastaanottajan tietämyksen tila, joka muuttuu tilannekohtaiseksi ongelma-avaruudeksi. Jos ongelma ei ratkea ajattelemalla, se johtaa epävarmuuden tilaan, jossa käytettävissä oleva tietämys ei ole riittävää ja lisätietämyksen hankinta informaatiota käytämällä on tarpeen.

Tuottajan hankkima tietämys ei ole sellaisenaan välitettävissä edelleen, vaan välitettävä informaatio tulee ensin tuottaa tekstinä l. esittää dokumenttina, joka välitetään. Toisaalta dokumentin sisällön ja muodon sanelee tietämys (se, mitä halutaan sanoa, sekä tietämys oletetuista vastaanottajista, kommunikaatiomuodoista ja tulkintamahdollisuuksista). Vastaanottaja tulkitsee dokumentteja ja luo niistä informaatiota, joka mahdollisesti auttaa epävarmuustilan voittamisessa. Dokumentista tulkittu informaatio ei välttämättä ole sama kuin tuottajan dokumenttina esittämä informaatio.



Kuva 1. Kognitiivinen kommunikaatiojärjestelmä (Ingwersen, 1992)

Informaatiosta

Informaation haun tutkimukseen soveltuvan informaatiokäsitteen tulee täyttää kahtalaiset vaatimukset : toisaalta informaatio on seuraus tuottajan tietämysrakenteiden tavoitteellisesta muokkauksesta (ottaen huomioon vastaanottajien tietämysten tilat) esitettynä dokumenttimuodossa, ja toisaalta se on jotakin, jonka havaitseminen vaikuttaa vastaanottajan tietämyksen tilaan ja muokkaa sitä (Ingwersen, 1992, 48). Tämän näkemyksen mukaan informaatio ei ole sama kuin teksti (tai muu dokumentin pintarakenne) vaan kätkeytyy siihen. Vastaanottaja löytää informaation, kun tulkitsee dokumentin. Dokumentti on kielen tasolla, informaatio kognitiivisella tasolla.

Seuraavassa lähdetään siitä, että kognitiivisella tasolla kommunikoitavaksi tarkoitettu tietämyksen tilan osa, informaatio, ajatellaan käsiterakenteeksi (tietämysrakenteeksi, käsitteelliseksi tietämysrakenteeksi, knowledge structure, ks. Ingwersen, 1992, 31–33). Se voidaan kuvata käsitteinä ja niiden välisinä suhteina. Kommunikoitaessa tämä käsiterakenne ilmaistaan tekstinä, se siis muotoillaan lingvistisellä tasolla luonnollisen kielen avulla (tai esim. kuvina).

Informaation hakujärjestelmät tallettavat ja käsittelevät dokumenttien ja/tai niiden kuvauksien digitaalisia esityksiä, eivätkä järjestelmät tulkitse niihin sisältyvää informaatiota. Järjestelmien sisältämät dokumentit (tai niiden kuvaukset) sinänsä eivät myöskään muuta kenenkään tietämysrakenteita. Näin ollen informaation hakujärjestelmät eivät sisällä todellista informaatiota vaan ainoastaan mahdollista informaatiota (potential information, Ingwersen, 1992, 48). Käsitellessään dokumenttien ja kysymysten esityksiä hakujärjestelmä kuitenkin käsittelee vain dataa, ei käsitteitä eikä luonnollista kieltä sinänsä vaan ainoastaan merkkijonojen esiintymiä. Ingwersenin mahdollinen informaatio on datan synonyymi. Tässä kirjoituksessa sanalla informaatio tarkoitetaan, kuten Ingwersenkin, mahdollista informaatiota, siis dataa.[2]

Käsite-, ilmaisu ja esitystasot

Informaatio siis on käsiterakenne, joka ilmaistaan kielen avulla dokumenttina, joka talletetaan datana hakujärjestelmään. Myös kysymys on käsiterakenne, joka ilmaistaan kielen avulla hakupyyntönä, joka esitetään kyselynä (datana) hakujärjestelmälle. Näistä näkökohdista seuraa, että kysymyksiä ja dokumentteja voidaan tarkastella kolmella tasolla: käsitetasolla, ilmaisutasolla, ja esitystasolla.

Käsitetasolla tarkastellaan kysymyksen ja dokumentin käsitteitä ja näiden suhteita. Hakuprosessissa tämä voidaan tehdä eksplisiittisesti sekä tallennusvaiheessa että hakuvaiheessa. Sekä indeksoinnin että haun oppaissa korostetaan käsiteanalyysiä välivaiheena ennen dokumentin tai hakupyynnön informaatiosisällön ilmaisemista esim. indeksitermeillä.

Ilmaisutasolla (lingvistisellä tasolla) tarkastellaan käsitteiden ilmaisutapoja luonnollisen kielen tai jonkin erikoiskielen (kuten dokumentaatiokielen) puitteissa. Luonnollista kieltä käytettäessä kuvataan kysymyksen ja dokumentin käsitteet ja näiden suhteet niitä edustavien luonnollisen kielen ilmausten (sanat, sanaliitot, fraasit, ym.) avulla. Dokumentin teksti on dokumentin täydellinen ilmaisutason esitys. Jos tietokannassa ei ole sovellettu dokumentaatiokieleen perustuvaa indeksointia, on hyvään hakutulokseen pääsemiseksi välttämätöntä kehittää myös kyselylle luonnolliseen kieleen perustuva esitys. Koska saman käsiterakenteen ilmaisut dokumentteina voivat olla hyvin monenlaisia, täytyy myös kyselyn ilmaisutason esityksen ottaa huomioon tämä moninaisuus – mahdollisia sanoja ja sanontoja löytyy usein monia. Jos luonnollisen kielen käsittelyn menetelmiä (lähinnä sanojen morfologinen analyysi ja lauseiden syntaksianalyysi) käytetään esitystasolla standardoimaan dokumenttien sanallisia esityksiä, ei ilmaisutasolla tarvitse pohtia kaikkea ilmaisujen moninaisuutta. Tällöin ei esim. tarvitse ottaa haussa huomioon sanojen taipumista.

Tietokoneisiin perustuva konkreettinen informaation haku tapahtuu aina esitystasolla. Kun haun suunnittelija pääsee esitystasolle, on käsitetaso ja ilmaisutaso käsitelty joko tietoisesti suunnitellen ja kehittäen tai alitajuisesti ja ongelmattomasti otaksuen. *Esitystasolla* käsitteellinen hakusuunnitelma saa vastineekseen hakuprofiilin, joka määrittelee hakujärjestelmälle, miten sen tulee toimia. Hakujärjestelmälle ominaisia näkökohtia dokumenttien ja kysymysten esityksessä ovat merkkijonojen samuus vs. erilaisuus (tai samuuden aste), merkkijonojen esiintymien keskinäinen sijainti (etäisyys), sijaintirakenne (esim. nimekekenttä) ja esiintymien lukumäärä. Esitystasolla kiinnitetään siis huomio

merkkijonojen esiintymiin hakuprofiileissa ja dokumenteissa, etenkin esiintymien keskinäiseen sijaintiin ja lukumääriin (tai muihin tilastollisiin ominaisuuksiin).

Koska dokumentteja ja kysymyksiä on tarpeen ja luontevaa tarkastella mainituilla kolmella tasolla, olisi suotavaa, että tarkastelussa käytettävät sanat myös heijastaisivat täsmällisesti kulloinkin tarkasteltavia tasoja.

4. Sana, termi ja käsite

Seuraavassa esitän tarkasteltavat merkkijonot, sanat, termit ja käsitteet eri lailla korostaakseni niiden erilaista luonnetta. Merkintä [X] (esim. [indeksitermi]) tarkoittaa käsitettä (merkitystä), jonka nimityksenä on X. Merkintä *X* (esim. *indeksitermi*) ilmaisee, että *X* on termi (se on jonkin käsitteen nimitys, mutta tässä sitä tarkastellaan terminä, siis kielen tasolla, eikä se ole käsite). Merkintä "X" (esim. "valtio") tarkoittaa luonnollisen kielen sanaa X sanana (sen merkitystä ei tällöin tarkastella). Merkintä X (esim. valt$) korostaa, että X on merkkijono – se voi olla, mutta sen ei tarvitse olla luonnollisen kielen sana. Voidaan todeta, että [X] kuuluu käsitetasolle, *X* ja "X" kuuluvat ilmaisutasolle ja X kuuluu esitystasolle. Sanan ja termin ero tulee seuraavassa esille.

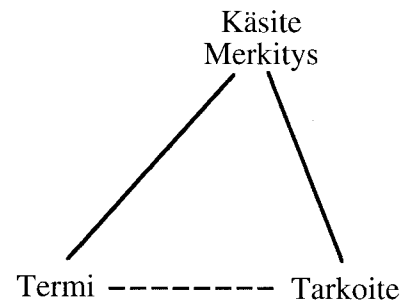
Yleis- ja erikoiskieli

Luonnollisen yleiskielen ja erikoiskielten ilmaisut rakentuvat sanoista. Sanat edustavat (nimeävät) käsitteitä, tosin kaikki sanat eivät edusta käsitteitä (esim. vieraiden kielten artikkelit "en", "the", jne.). Luonnollisen kielen sanojen nimeämät käsitteet ovat kuitenkin kussakin kommunikaatioyhteisössä vain osittain yhteisiä – sanan edustama käsite riippuu kielen tulkitsijasta. Eri tulkitsijoiden nimeämät käsitteet [demokratia], [koulutus] tai [pyyhekumi] eivät ole täysin samoja. Niiden käsitepiirteet ja merkitys ovat osittain yhteisiä, sillä muutoin toinen toisensa ymmärtäminen ja mielekäs kommunikaatio ei olisi mahdollista. Yhteisyys on kuitenkin vain osittaista, sillä jokaisella tulkitsijalla on omasta taustastaan (kokemus, tavoitteet, pätevyys) johtuvia erikoispiirteitä käsitteiden määrittelyssä. Tarkkaan ottaen henkilöiden A ja B käsitteet [demokratia$_A$] ja [demokratia$_B$] eivät aina ole sama käsite, vaikka he käyttävät käsitteistään samaa nimitystä "demokratia".

Konkreettisten ja yksinkertaisten käsitteiden (esineiden, tapahtumien, tekojen) käsitepiirteet ovat yleisesti tunnettuja ja hyväksyttyjä (esim. [pyyhekumi], [auton rengas]). Henkilökohtaisen tulkintavaran osuus on suhteellisen pieni. Abstraktien ja monimutkaisten käsitteiden käsitepiirteet ovat vain osittain yleisesti tunnettuja ja/tai hyväksyttyjä (esim. [demokratia] tai [koulutus]). Henkilökohtaisen tulkintavaran osuus on näissä suhteellisen suuri. Tämän takia esim. tutkimusraporteissa tulee pyrkiä täsmentämään käytettävät käsitteet. Lukija saa paremman mahdollisuuden ymmärtää raportti kirjoittajan tarkoittamalla tavalla.

Sanastotyön (Haarala, 1981) ja käsiteanalyysin (Niiniluoto, 1980) piirissä on sanalla "termi" rajattu merkitys : Termi on erikoiskielen ilmaisu, jolla on ko. erikoiskielessä täsmennetty (rajattu) ja yleisesti hyväksytty merkitys. Usein termin merkitys leikkaa vastaavan luonnollisen kielen ilmaisun merkitystä, mutta rajoittaa ja täsmentää sitä tehden merkityksen yksikäsitteiseksi. Termi nimeää käsitteen, jolla on jokin (usein kielen ulkoinen) tarkoite (referentti). Termin, käsitteen ja tarkoitteen suhdetta voidaan havainnollistaa Ogdenin ja Richardsin kolmion avulla (kuva 2, ks. myös Niiniluoto, 1980, 118).

Näin ollen erikoiskielen termi ei ole sama kuin jokin luonnollisen kielen sana, vaikka ne kirjoitettaisiinkin samalla tavalla. Esim. sanan "saanti" esiintyminen tekstissä saattaa liittyä kalastukseen



Kuva 2. Termin, käsitteen ja tarkoitteen suhde (Haarala, 1981)

tai kuvaputkien valmistukseen. Mutta jos sanan "saanti" sisältävä teksti kuuluu informaation haun erikoiskielen piiriin, on sana "saanti" useimmiten erikoiskielen termi *saanti,* jolla on rajattu merkitys [saanti]. Viitetiedostosta saattaa löytyä viite, jossa esiintyy indeksitermikentässä (indeksi)termi *luukato* ja tiivistelmän tekstissä sana "luukato". Indeksitermi *luukato* on termi, indeksoinnin erikoiskielen ilmaisu. Tiivistelmän sana "luukato" on joko tavallinen sana tai kuuluu lääketieteen erikoiskieleen ja on siis sittenkin termi. Tämä riippuu kyseisen dokumentin luonteesta (kuulumisesta lääketieteen erikoiskieleen, jos tässä on termi *luukato*).

Tuottaja, indeksoija, asiakas, välittäjä, dokumentti ja tietokanta tulkinnan kontekstina

Jos käsitteen [termi] määritelmästä ([termi] on erikoiskielen käsitteen yleisesti hyväksytty ja yksikäsitteinen nimitys) pidetään vakavasti kiinni, muodostaa koneellinen informaation haku tarkastelukehyksenä ongelman. Jotta jotakin sanaa voitaisiin haussa pitää terminä, asettaisin seuraavan, termin määrittelyyn perustuvan täsmentävän ehdon: sanan (merkkijonon) tulee olla termi kommunikaatioprosessin kaikissa osissa, joissa sitä käytetään. Toisin sanoen jokaisen dokumentin tuottajan tulee hallita termi jokaisessa tietokantaan tallennetussa dokumentissa ja jokaisen indeksoijan tulee hallita se indeksoinnissaan. Lisäksi tarkasteltavan asiakkaan ja häntä avustavan välittäjän tulee hallita termi omassa kommunikaatiossaan. Kukaan ei saa pettää ketjussa.

Seuraavassa tarkastelen vain sellaisia termejä, joita edustavat merkkijonot ovat myös luonnollisen kielen sanoja – luokitusnotaatiot yms. jätän tarkastelun ulkopuolelle. Ilmauksella 'X ei tunnista sanaa "Y" termiksi' tarkoitan seuraavassa, että tulkitsija X ei tiedä, kuuluuko tarkasteltava sana "Y" johonkin erikoiskieleen vai ei. Se saattaa sellaiseen kuulua tai olla kuulumatta. Tulkitsijalla X ei ole pätevyyttä ko. erikoiskielessä, jos se on olemassa. Tulkitsija X ei välttämättä tiedä, onko erikoiskieltä olemassa. Seuraavien viiden kappaleen lopussa suluissa olevat kommentit ovat kärjistyksiä.

Jos dokumentin tuottaja ei tunnista käyttämäänsä sanaa "Y" termiksi, kyseinen sana ei ole termi hänen dokumentissaan. Esim. tämän kirjoittaja ei tiedä, onko "luukato" lääketieteen termi vai ei. Tuottajan tulee tuntea myös käsite [Y], jotta hänen dokumentissaan voisi esiintyä *Y.* Ilman näiden ehtojen täyttymistä merkkijonon Y esiintymät dokumentissa pitää ajatella sanan "Y" esiintymiksi, ei termin *Y* esiintymiksi. Sanan "Y" edustama käsite jää avoimeksi. (Puoskareiden dokumentit kuuluvat siis lähinnä sanataiteen alaan).

Jos tietokannassa tai sen tarkasteltavassa osassa on yksikin dokumentti, jossa esiintyvä sana "Y" ei samalla ole termi *Y,* ei tietokannassa tai sen tarkasteltavassa osassa ole termiä *Y.* Haku merkkijonolla Y ei tuota yksikäsitteistä tulosta. (Puoskareiden dokumentteja ei siis pitäisi tallentaa tietokantaan).

Oletetaan että tietokannan (sen tarkasteltavan osan) dokumentit sisältävät termin Y. Jos indeksoija ei tunnista dokumentista lukemaansa olennaista sanaa "Y" termiksi, kyseinen sana ei ole termi hänen muodostamassaan kuvassa dokumentista. Jos hän ei löydä sanaa "Y" dokumentaatiokielestä (viittausterminä tai indeksiterminä), jää kirjoittajan [Y] helposti indeksoimatta tai eri indeksoijat indeksoivat sen varsin eri tavoin. Tämän todennäköisyys kasvaa, jos indeksoija ei tunne käsitettä [Y]. Sana "Y" voi olla sekä termi dokumentin erikoiskielessä että indeksitermi dokumentaatiokielessä. Käytän näistä termeistä merkintöjä Y_d ja Y_t. Jos indeksoija indeksoi dokumentin termillä Y_t, mutta ei tunne käsitettä [Y_t], indeksointituloksessa ei ole termiä Y_t, vaan sana "Y_t". Termin yksikäsitteisyysvaatimuksesta seuraa, että jos yksikin indeksoija käyttää dokumentaatiokielen termiä Y_t, mutta ei tunne käsitettä [Y_t], niin tietokannassa tai sen tarkasteltavassa osassa ei esiinny termiä Y_t, vaan vain sana "Y_t". (Ei siis pidä käyttää pätemättömiä indeksoijia).

Oletetaan, että tietokanta (sen tarkasteltava osa), indeksointi mukaan lukien, sisältää termin *Y.* Jos välittäjä ei tunnista käyttämäänsä sanaa "Y" termiksi, kyseinen sana ei ole termi hänen hakuprofiilissaan. Jos hän ei tiedä sitä dokumentaatiokielen indeksitermiksi, hän ei voi käyttää sitä indeksiterminä, vaan joutuu hakemaan sitä kaikista tekstikentistä. Jos hän ei tiedä sitä dokumenttien erikoiskielen termiksi, hän joutuu miettimään sanalle "Y" vaihtoehtoisia ilmauksia. Näyttää kuitenkin siltä, että välittäjän hakusuunnitelmassa voi olla *Y,* jos hän tietää "Y":n termiksi, vaikka ei tuntisi käsitettä [Y]. Riittääkö termin esiintymiseen hakusuunnitelmassa vielä sekin, että asiakas tunnistaa sanan termiksi ja välittäjä uskoo tämän? Ilmeisesti riittää. Keskustelun tuloksena voidaan ajatella välittäjän oppivan tunnistamaan termin *Y,* vaikka ei opikaan

käsitettä [Y]. Tilanne lienee tavallinen informaatiopalvelussa. (Jotain toivoa informaatikollakin ...).

Oletetaan edelleen, että tietokanta (sen tarkasteltava osa), indeksointi mukaan lukien, sisältää termin *Y*. Jos asiakas (informaation tarvitsija) ei tunne sanaa "Y" termiksi, kyseinen sana ei ole termi hänen kysymyksessään. Helposti asiakas yrittää kuvailla informaatiotarvettaan myös muilla sanoilla, jotka tarjoavat välittäjälle (hakujärjestelmälle) lisää ymmärtämis- tai harhautumismahdollisuuksia. Jos esim. meidän muori on kiinnostunut aiheesta luukato, on kyseessä useimmiten vain "luukato", "se kun reisi murtuu liukastellessa" eikä lääketieteen *luukato*. Muorin [luukato] ja häntä hoitavan lääkärin [luukato] lienevät löyhässä assosiaatiosuhteessa. Riittääkö muorin haussa, jos hän tietää, että kyseessä on *luukato* (sehän saattoi lukea esim. lääkärin diagnoosissa), mutta hänelle on epäselvää, mikä on [luukato] ? Näyttää siltä, että asiakkaan kysymyksessä voi olla *Y*, vaikka asiakas ei tuntisi käsitettä [Y]. (Asiakkaan ei sentään tarvitse olla pätevä).

Jos kaikki tuntevat "Y":n Y:ksi, niin termiperusteinen kommunikaatio, vanha unelma, voi toteutua. Mutta ankarasti ottaen sana "Y" ei ole termi informaation haun kontekstissa, jos joku keskustelijoista (tuottaja, indeksoija, asiakas ja välittäjä) ei tunne sanaa "Y" termiksi (ja jos tuottaja ja indeksoija eivät tunne käsitettä [Y]). Vaatimukseni onkin mahdoton, eikä kaiketi tavallisesti toteudu hakutilanteissa. Relevanssiongelmat ovatkin tyypillisiä informaation haussa, eikä syyttä. Toisaalta on selvää, että monta aihepiiriä kattavissa tietokannoissa, joissa on tekstikenttiä (esim. tiivistelmiä), ei tekstikentissä koko tietokannan kontekstissa esiinny termejä. Jos jossakin haussa kyetään rajaamaan tietokannasta sellainen osa, jossa esiintyy vain tietyn erikoiskielen puitteissa kirjoitettuja dokumentteja ja asiakas ja välittäjä tuntevat termit (ei välttämättä niiden merkityksiä), on haussa mahdollisuus käyttää termejä.

Vaikka sanastotyötä tehdäänkin laajalti, harvalla tiedonalalla lienee erikoiskieltä, jonka termien puitteissa dokumentit kirjoitettaisiin. Usein yksikäsitteisyysehto ei aivan toteudu ja käytettävissä on muutamia synonyymejä, lähes-termejä (luukato, osteoporoosi?). Kussakin dokumentissa voidaan teknisin määritelmin luoda dokumentin oma erikoiskieli, mutta se on voimassa vain ko. dokumentissa. "Hakutermi" ei tässä tilanteessa ole hyvä nimitys, koska sen viittaa käsitteeseen [termi], jonka käytön edellytykset eivät ole voimassa.

Jos haku perustuu jonkin dokumentaatiokielen indeksitermeihin, on luonnollista kutsua indeksitermejä termeiksi myös ko. haussa – muun vaatiminen, hyvät pyrkimykset ja keskustelun sujuvuus huomioon ottaen, olisi kai kohtuutonta. Myös nimitys "indeksitermi" olisi johdonmukainen nimitys hakujen muodostamisessa.

Mielestäni ei ole syytä luopua käsitteen [termi] määrittelytavasta, koska se on hyvin juurtunut niin sanastotyöhön kuin dokumentaatiokieltenkin tarkasteluun. Näin ollen nimitystä termi pitäisi käyttää vain rajatuissa merkityksissä. Indeksitermeihin perustuvassa haussa käytettäisiin (haku)termejä ja muulloin (haku)sanoja paitsi joissakin erikoistilanteissa, joissa muun erikoiskielen termien käytettävyys termeinä on taattu. Seuraava ehdotukseni perustuu tähän näkemykseen.

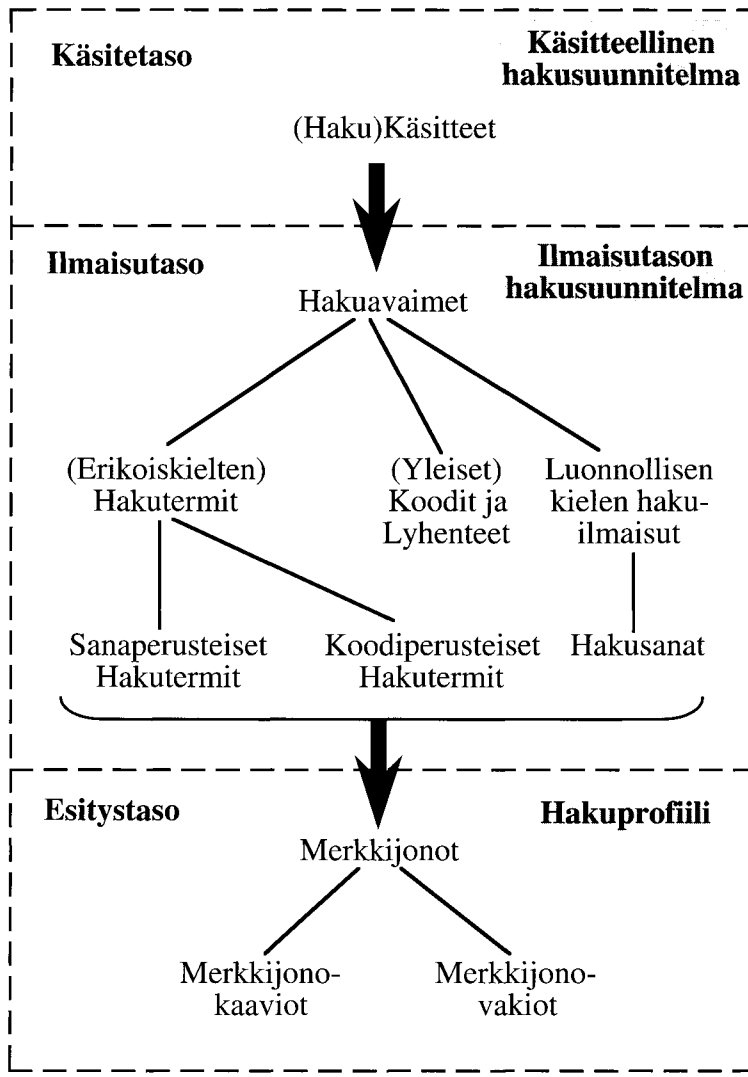
5. Käsitteet, termit, sanat ja merkkijonot haun kannalta

Luvussa 2 todettiin, että dokumentteja ja kysymyksiä voidaan tarkastella käsite-, ilmaisu- ja esitystasoilla, minkä vuoksi olisi suotavaa, että tarkastelussa käytettävät termit myös heijastaisivat täsmällisesti niitä tasoja, joita kulloinkin käytetään. Informaation haku suunnitellaan periaatteessa tasoperiaatteen mukaisesti siten, että aluksi analysoidaan (käsitetasolla) hakupyynnön käsitteet ja niiden suhteet käsitteelliseksi hakusuunnitelmaksi, seuraavaksi selvitetään, miten nämä käsitteet voidaan ilmaista ilmaisutason hakusuunnitelmana (ilmaisutasolla) ja lopuksi viimeistellään (esitystasolla) hakuprofiili, jossa tarkastellaan kirjoitusasuja, katkaisua ja läheisyysoperaatioita (siis merkkijonojen esiintymistä). Luvun 3 tuloksena oli, että informaation haun muodostamassa kommunikaatiossa on ankarasti ottaen vähän termejä, mutta paljon sanoja. Armoa annettiin kuitenkin dokumentaatiokielten (indeksi)termeille. Tämä rajankäynti koski ilmaisutasoa. Nämä seikat huomioonottaen ehdotan seuraavaa terminologiaa informaation hakua varten (kuva 3).

Käsitetasolla tarkastellaan dokumentteihin ja kysymyksiin sisältyviä (haku)käsitteitä ja niiden muodostamia rakenteita. Hakukäsitteitä ei voi suoraan poimia dokumenteista tai kysymyksistä, vaan ne täytyy tulkita niiden ilmaisutason esityksistä. Käsitetasolla ei ole sanoja tai termejä, vaan yksiköt ovat käsitteitä (kuva 3).

Ilmaisutasolla on hakuavaimia, jotka ovat hakutermejä, luonnollisen kielen ilmaisuja tai koodeja ja lyhenteitä.

- *Hakuavain* on myös yleisnimitys luonnollisille hakusanoille, yleiskieleen sisältyville lyhenteille (esim. "FIM") ja koodeille (esim. "20-25°") sekä hakutermeille silloin, kun erottelulla näihin luokkiin ei ole merkitystä.
- *Hakutermi* on dokumentaatiokielen tai muun tietyn erikoiskielen termi, jota käytetään käsitteellisen hakusuunnitelman sisältämän hakukäsitteen nimityksenä ilmaisutason hakusuunnitelmassa. Jos hakutermi on dokumentaatiokielen termi, tulee sen esiintymiä etsiä dokumenttien sellaisista kentistä, jotka on varattu dokumentaatiokielen termejä varten (usein nimeltään indeksitermikenttä). Jos hakutermi on jonkin muun erikoiskielen termi, tulee sen esiintymiä etsiä dokumenttien sellaisista kentistä, jotka on varattu ko. kielen termejä varten (esim. tekstikenttä). Hakutermi voi olla sanaperusteinen (sen sisältämät sanat kuuluvat myös luonnolliseen kieleen, esim. *tietoliikenneprotokolla*) tai koodi-



Kuva 3. Merkkijonot, hakuavaimet ja käsitteet haun kannalta

perusteinen, jolloin se ei esiinny luonnollisessa kielessä, kuten esim. *udk:681.3-051, X.500 -protokolla.* Hakutermi voi koostua monesta sanasta tai yhdyssanasta.

• *Luonnollisen kielen hakuilmaisut* ovat luonnollisen kielen yksittäisiä sanoja ja yhdyssanoja (yhdessä *hakusanoja,* kuten "nopeusmittari", "juna") ja niistä rakentuvia monimutkaisempia luonnollisen kielen ilmaisuja (esim. "lakisääteinen eläke- ja vahinkovakuutus", "lukemisrajoitteisten opetusmateriaali").

Hakuavaimet esiintyvät tiedostojen tietueissa useimmiten vaihtelevissa kieliopillisissa muodoissa. Esitystasolla hakuavainten esiintymiä tarkastellaan merkkijonoina, joiden perusominaisuuksia on samuus (vs. erilaisuus), sijainti tietueissa ja esiintymien lukumäärä. *Merkkijonot* ovat käytettävän kielen merkeistä koostuvia, erotinsymbolien rajaamia jonoja. Hakujärjestelmät toimivat vain esitystasolla, vaikka käsiteltävät dokumenttien esitysten osat, merkkijonot, voivatkin olla käsiteindeksoinnin tuloksia (jolloin käsitellään hakutermien esiintymistä) tai luonnollisen kielen hakuilmaisuja (jolloin käsitellään sanojen tai fraasien esiintymistä). Jos dokumenttien tallennuksessa on palautettu sananmuodot kieliopillisiin perusmuotoihinsa, voi hakujärjestelmä vertailla näitä perusmuotoja edustavien merkkijonojen samuutta, sijaintia ja esiintymien lukumääriä. Hakujärjestelmän käsittelemät merkkijonot ovat joko merkkijonokaavioita tai merkkijonovakioita. *Merkkijonokaaviot* täsmäävät useisiin merkkijonovakioihin. *Merkkijonovakiot* edustavat esitystasolla ilmaisutason hakusanoja. *Merkkijonon (hakusanan) katkaisu* (truncation) lopusta on merkkijonokaavioiden tavallisin erikoistapaus.

Periaatteessa esitystasolla ei pitäisi puhua muista kuin merkkijonoista, koska hakujärjestelmät käsittelevät vain sellaisia. Vaikka jokin merkkijonovakio olisi täsmälleen jonkin luonnollisen hakusanan esitysmuoto luonnollisessa kielessä, on se hakujärjestelmän kannalta vain merkkijono – siihen ei periaatteessa pidä liittää niitä kieleen kuuluvia lisäpiirteitä, joita se saa, jos sitä kutsutaan hakusanaksi. Merkkijonovakio hakuprofiilissa on hakujärjestelmälle vain merkkijono eikä merkitse yhtään mitään. Vielä vaikeampaa on ajatella merkkijonokaaviolle mitään merkitystä.

Käytännössä merkkijonovakioiden (esitystaso) ja hakuavainten (ilmaisutaso) läheinen suhde johtaa puheen yksinkertaistamiseen: myös esitystasolla puhutaan mm. hakusanoista. Koska merkkijonokaaviot täsmäävät useisiin merkkijonovakioihin, voidaan yksinkertaistaen puhua myös hakuavainkaavioista, joita täsmäytetään kaikentyyppisiin hakuavaimiin, hakutermikaavioista, joita täsmäytetään vain hakutermeihin, ja hakusanakaavioista, joita täsmäytetään vain hakusanoihin – näin voitaisiin jatkaa myös koodikaavioihin ja lyhennekaavioihin. Tällainen kielenkäytön yksinkertaistaminen ei ole kovin vaarallista kunhan muistetaan se, että hakujärjestelmälle sanat ja termit ja niihin liittyvät kaaviot ovat vain merkkijonoja.

Kuvan 3 nuolet kuvaavat sitä, että tasojen välillä tehdään käännöksiä – kuvassa aina kohti esitystasoa, mutta periaatteessa (dokumenttia tai kysymystä tulkittaessa) myös toiseen suuntaan. Ylempi nuoli ilmaisee, että käsitetason käsitteet käännetään ilmaisutason hakuavaimiksi. Jotakin käsitettä voi edustaa hakusana, toista lyhenne ja kolmatta kenties hakutermi. Alempi nuoli ilmaisee, että kaikki hakuavaimet käännetään hakujärjestelmää varten hakuprofiilin merkkijonoiksi (vakioiksi tai kaavioiksi).

Nimityksiä "kontrolloitu termi", "kontrolloimaton termi" ja "vapaa termi" ei pitäisi käyttää lainkaan. Nimitys "kontrolloitu termi" viittaa dokumentaatiokielen tai muun erikoiskielen termiin ja niihin sisältyy jo tämän takia rajattu merkitys ("kontrollointi"), jota ei tarvitse toistaa. Nimitys on siis täysin turha. Nimitykset "kontrolloimaton termi" ja "vapaa termi" taas viittaavat erikoiskieliin kuulumattomiin luonnollisiin sanoihin, joihin ei pidä soveltaa nimitystä termi lainkaan. Nämä nimitykset ovat sisäisesti ristiriitaisia. Nimitys vapaatekstihaku tarkoittaa joskus hakua luonnollisilla sanoilla ja joskus hakua dokumentin tekstikentästä (≈ "kokotekstihaku"). Edellisessä tapauksessa parempi ilmaisu on (haku)sanahaku. Jälkimmäisessä tapauksessa suosisin ilmausta tekstihaku erotuksena indeksitermihausta, nimekehausta tai tiivistelmähausta.

6. Keskustelu ja johtopäätökset

Ehdotukseni termeiksi käsite, hakuavain ja merkkijono alatermeineen ja merkityksineen on ehdotus. Termeiksi tuleminen edellyttäisi, että ne olisivat yleisesti hyväksyttyjä ja käytettyjä sekä yksikäsitteisiä käsitteiden nimityksiä. Kuitenkin tässä vaiheessa tarvitaan keskustelua sekä niiden käsitteiden määrittelystä, joita termiehdotukseni

nimeävät, että myös termien muodosta. Ehdotukseni hyvinä puolina pidän seuraavia :

- termien nimeämät käsitteet voidaan määritellä riittävän tarkkaan
- termit eivät ole outoja tai kovin kaukana informaation haun keskustelusta – ne voidaan omaksua helpohkosti, koska sanoina niillä jo on lähes oikea merkitys
- termit ovat varsin yksinkertaisia eivätkä turhia tai ristiriitaisia
- monia muita keskustelussa esiintyviä sanoja voidaan jättää pois käytöstä
- niiden avulla voidaan tehdä hakujen tarkastelutasojen mukaiset erottelut.

Termien muotoa kannattaa miettiä myös siltä kannalta, kuinka termit luontuvat puheeseen. Esim. *hakutermi* sinänsä tuntuu selkeältä, mutta hakutermihaku ei ole luonteva. Olisiko *indeksitermi* parempi muotoilu? Indeksitermihaku on luontuva nimitys. Nimitys *indeksitermi* myös viittaa eksplisiittisesti siihen, että termi kuuluu dokumentaatiokieleen. Onko *hakukäsite* parempi nimitys kuin lyhyempi *käsite* ? Indeksoinnissa tulee silloin kenties tarve termille *tallennuskäsite.*

Dokumentaatiokieliin ja informaation tallennukseen liittyvää terminologiaa en ole varsinaisesti tarkastellut. Kielenkäyttö informaation tallennuksen alueella tarjoaa saman rikkauden ja sanataiteen mahdollisuudet kuin haunkin alueella. Ajateltakoon seuraavia: asiasana, avainsana, indeksitermi, kuvailutermi, suositeltu termi, deskriptori, vapaa termi, vapaa indeksitermi, kontrolloitu indeksitermi, kontrolloimaton indeksitermi, hakemistotermi, suppeampi termi, laajempi termi, rinnakkaistermi, jne. Voitaisiinko kuvan 3 jäsennystä hyödyntää? Hakutermi korvattaisiin indeksitermillä ja sen alatermejä olisivat hakemistotermi, suppeampi termi, laajempi termi, rinnakkaistermi? Tallennuksessa ei tarvita termiä merkkijonokaavio (vain vakiot tallennetaan).

Sanan käytön, termin tuntemuksen ja käsitteen hallinnan sekä toisiin keskustelijoihin kohdistuvan luottamuksen (tai uskon tai sen puutteen) erittelyä (luku 3) voi jatkaa pidemmälle ja se johtaa varsin tuttuja kysymyksiä koskeviin, mutta mielenkiintoisiin tarkasteluihin.

Viitteet

1) Käytän tässä kirjoituksessa nimitystä *informaation tallennus ja haku* siinä merkityksessä kuin Peter Ingwersen (1992) sitä käyttää. Useimmat lukijat käyttäisivät suomennosta *tiedon tallennus ja haku.* Olkoon niin.
2) Miksi sitä pitää sitten kutsua (mahdolliseksi) informaatioksi? Datan haku (data retrieval) tarkoittaa useimman puheessa hakua rakenteisista tiedostoista, jolloin vastaukset koostuvat faktoista, ei teksteistä. Esim. SQL-kysely (Date, 1986) Hämeen läänin työttömyydestä työttömyystietoja sisältävään tiedostoon voisi olla seuraavanlainen :

```
select   kunta, työttömyysprosentti
from     TYÖTTÖMYYSTIEDOSTO
where    lääni = "Häme"
```

Kyselyn vastaus olisi taulukko, jonka ensimmäinen sarake sisältäisi kuntien nimiä ja toinen niiden työttömyysprosentteja. Jollakin rivillä saattaisi lukea "Tampere, 18", joka tulkitaan niin, että Tampereen työttömyysprosentti tiedoston kattamana aikana (hetkenä) on 18 %. Tällaista hakua moni kutsuu datan hauksi. Informaation haku taas tuottaisi vastauksenaan tekstejä. Jos ajatellaan, että todellinen informaatio on se, mikä luodaan mahdollisen informaation (datan) tulkinnan avulla, on vaikea tehdä eroa datan haun ja informaation haun välillä. Molemmissa on kyse datan tulkinnasta, jolloin saadaan informaatiota, joka voi muuttaa tulkitsijan tietämystä.

Lähteet

Date, C.J. (1986). An introduction to data-base system. Vol 1, 4th ed. Reading, MA : Addison-Wesley.

Haarala, R. (1981). Sanastotyön opas. Helsinki : Tekniikan Sanastokeskus.

Ingwersen, P. (1992). Information retrieval interaction. London, UK: Taylor Graham.

Niiniluoto, I. (1980). Johdatus tieteenfilosofiaan : Käsitteen- ja teorianmuodostus. Helsinki : Otava.

Hyväksytty julkaistavaksi 28.11.1993