

Janne Heinonen

TESAURUKSET JA ONTOLOGIAT

TJTSD50 - Tekstiedonhaku
Esseetehtävä
23.5.2006

Jyväskylän yliopisto
Tietojenkäsittelytieteiden laitos
Jyväskylä

SISÄLLYSLUETTELO

1 TESAURUKSET JA ONTOLOGIAT.....	3
1.1 Johdanto	3
1.2 Tesaurukset.....	4
1.3 Ontologiat	6
1.4 Yhteenveto	9
LÄHDELUETTELO	10

1 TESAURUKSET JA ONTOLOGIAT

1.1 Johdanto

Perinteiset tesaurokset on laadittu ihmisten tarpeita silmällä pitäen ja termien merkitysten kuvaus on voitu jättää ihmisen tulkinnan varaan. Yhä enenevässä määrin tesauroksia käyttävät myös tietokonesovellukset, joille sanastot tulisi aiempaa täsmällisemmin määritellä ontologioina. Koneelle yksittäisellä termillä ei ole merkitystä, vaan merkitys syntyy termien välisten suhteiden kautta, esimerkiksi termi "Suomi" on hierarkkisessa osa-kokonaisuussuhteessa termiin "Pohjoismaat". Suhteiden täsmällinen kuvaaminen on siten koneiden kannalta erityisen tärkeää. Tieto- ja tietämystekniikan näkökulmasta tarkasteltuna tesauroksetkin ovat eräänlaisia ontologisia kuvauksia maailmasta ja niitä voidaan käyttää hyväksi tiedon haussa, päättelyssä ja esittämisessä monin tavoin (Hyvönen 2005).

Sekä ontologiat että tesaurokset yhdistetään käsitteeseen taksonomia. Taksonomia on järjestelmä, joka soveltuu erityisesti informaatio-olioiden semanttiseen luokitteluun. (Daconta, Obrst & Smith 2003) Hyvä käytännön esimerkki taksonomiasta luokkineen ja alaluokkineen on kirjojen järjestys kirjastoissa tai vaihtoehtoisesti vaikkapa puhelinluettelosta löytyvät kategorisoidut keltaiset sivut.

Edellä kuvatut terminologiset määrittelyt liittyvät läheisesti myös semanttisen tietoverkon ideaan. Semanttisen webin kehitystyössä pohditaan miten saada nykyinen tiedon visualisointiin keskittyvä World Wide Web älykkäämmäksi lisäämällä siihen eritasoisia sisällöllisiä kuvauksia, semantiikkaa. (Hyvönen 2001)

1.2 Tesaurukset

Tesaurus-sanan juuret juontavat kreikan ja latinan kielistä, joissa sanaa on käytetty tarkoittamaan ”sanojen aarreaittaa”. Nykykäsityksen mukainen tesaurus rakentuu yksinkertaisimmillaan valmiiksi koostetusta listasta, joka sisältää kohdealueen tärkeimmän sanaston. Lisäksi jokaiseen sanaan liittyy joukko siihen suhteessa olevia sanoja. Tesaurukset pitävät yleisesti yksittäisten sanojen lisäksi sisällään myös laajempia ilmaisuja, kuten fraaseja. (Baeza-Yates & Ribeiro-Neto 1999, 170) Järvelinin (1995, 150) mukaan tesaurusten suhdeverkostot rakentuvat assosiaatiosuhdetyyppien varaan, joita voivat olla tyypillisesti esimerkiksi:

- käsitteelliset veljekset, joiden tarkoitteet leikkaavat tai mielletään samantapaisiksi (talvirengas, nastarengas, vyörengas)
- syy – seuraus (alkoholismi – juopottelu)
- kohde – toiminto (kirjasto – lainaus)
- prosessi – liitännäisprosessi (johtaminen – päätöksenteko)
- käsitteet joiden käsitepiirteiden yhtäläisyys on suuri (myrkkyy, myrkyllisyys)

Daconta ym. (2003) luettelevat tesauruksista löytyviksi suhdetyypeiksi ekvivalenssin, homografian, hierarkian ja assosiativisuuden. Ekvivalenssi tarkoittaa samankaltaisuutta ja synonyymisuhdetta, homografiassa sanoilla on sama kirjoitusasu, mutta eri merkitys, hierarkiassa toinen on kapeampi tai laveampi käsite kuin toinen ja assosiativisuus ilmenee esimerkiksi sanaparissa naula – vasara. Defuden (1984) mukaan käsitteille voidaan laatia myös numeerisia arvoja semanttisen etäisyyden suhteen, kuten esimerkiksi täyttä vastaavuutta (1) hieman heikompi arvo 0.8 englanninkieliselle sanaparille boat – ship.

Tesaurusten indeksikomponentit ovat termejä, jotka yleensä pitävät sisällään tietyn käsitteen. Käsite on puolestaan semanttisen tiedon perusyksikkö, jonka avulla ilmaistaan ideoita. Termit ovat useimmiten substantiiveja, sillä ne muo-

dostavat konkreettisimman osan kielestä. Usein käytetään monikkomuotoa, sillä tesaureissa pyritään ilmaisemaan asioiden luokkia, esimerkiksi luokka: ohjukset – alaluokka: ballistiset ohjukset. (Baeza-Yates & Ribeiro-Neto 1999, 171) Yleisessä suomalaisessa asiasanastossa (Helsingin yliopiston kirjasto 2006) yksikkömuodossa ovat esimerkiksi ainesanat sekä abstrakteja käsitteitä ja toimintaa kuvaavat sanat. Monikkomuotoisia ovat yleensä konkreettiset, laskettavissa olevat käsitteitä kuvaavat sanat.

Tesaurusta voidaan käyttää asiasanoituksen ja luokituksen (indeksoinnin) apuvälineenä, tiedon haun apuvälineenä tai molemmissa tehtävissä (Hyvönen 2005). Tesaurus mahdollistaa yhtenäisen sanaston käyttämisen indeksoinnissa ja hakutehtävissä, avustaa asianmukaisten hakutermien löytämisessä sekä tarjoaa luokitellun hierarkian, jonka avulla on helppo laajentaa tai supistaa hakuja. (Baeza-Yates & Ribeiro-Neto 1999, 170) Sanaston joihinkin asiasanoihin voi liittyä myös selityksiä. Selityksissä annetaan mm. ohjeita sanastoon sisällytettävien asiasanojen käytöstä ja täsmennetään joidenkin asiasanojen merkitystä.

Tesaurukset tarjoavat hyödyllistä apua kyselymuodostuksen apuvälineenä. Kun tiedonhakija aloittaa prosessinsa, hänen täytyy muodostaa ensin käsitteistö hakemastaan asiasta. Tämä hakutehtävään liittyvä informaatiotarve täytyisi kyetä eksplikoimaan hakulausekkeeksi tietokantaan. Tässä kohtaa tesaurus voi tarjota apuaan, mutta toisaalta tesaurus voi johtaa myös harhaan - Tesaurus on voitu laatia jonkin muun henkilön toimesta ja eri tarvetta ja kontekstia ajatellen. (Baeza-Yates & Ribeiro-Neto 1999, 172)

1.3 Ontologiat

Alkujaan sana ontologia on kuulunut filosofian piiriin, missä sillä on käsitetty olevaisen tutkimista. Nykyisin termi mielletään kuitenkin mieluummin IT-alan sanastoon kuuluvaksi. Ontologiat ovat formaaleja eksplisiittisiä määrittelyitä yhteisestä käsitteistöstä, jotka mahdollistavat myös käsitteistön koneellisen tulkinnan. Käsitteistön yhteisyys mahdollistaa tietämyksen jakamisen, yhteiskäytön ja yhdistämisen. (Gruninger & Lee 2002) Ontologian luomisen keskeiset välineet ovat ontologiakieli, jolla käsitteet ja niiden väliset suhteet määritellään sekä ontologiaeditori, joilla ontologiset kuvaukset käytännössä laaditaan. (Hyvönen 2001)

Ontologian avulla voidaan esittää jonkin erityisalan ammattikäsitteitä ja -tietämystä, metadataa, yleistä arkitietämystä, käsitteistöjä, tehtäviä sekä prosesseja ja palveluita. Metadataa voisivat olla esimerkiksi tietolähteen tai kuvan julkaisutiedot. Tunnettuja laajoja ontologioita ovat mm. WordNet (<http://www.cogsci.princeton.edu/~wn/>), joka sisältää yli 100.000 englannin kielen käsitettä ja IT- ja elektroniikkateollisuuden RosettaNet (<http://www.rosettanet.org>). (Hyvönen 2001)

Dacontan ym. (2003) mukaan ontologiaan sisältyvät:

- Yleiset luokat
"Classes (general things) in the many domains of interest"
- Luokkien instanssit
"Instances (particular things)"
- Suhteet edellisten välillä
"Relationships among those things"
- Ominaisuudet ja ominaisuuksien arvot
"Properties (and property values) of those things"
- Toiminnot ja prosessit
"Functions of and processes involving those things"

- Rajoitteet ja säännöt
"Constraints on and rules involving those things"

Ontologioiden peruskäsitteisiin kuuluu siis luokka, johon sisältyy alaluokkia. Luokat ovat luokkahierarkiassa, jonka alimmalla tasolla ovat ilmentymät. Ilmentymät ovat kyseessä olevaan luokkaan liittyviä yksilöitä eli jäseniä. (Chandrasekaran, Josephson & Benjamins 1999) Tämä on lähellä olioparadigman mukaista lähestymistapaa, jossa ominaisuudet periytyvät alaluokille, ja missä ilmentymät pitävät sisällään arvoja ominaisuuksille.

Hendlerin (2001) mukaan ontologioihin kuuluvat myös päättely- ja logiikkasäännöt. Ontologia on tällöin joukko tietämystermejä, joka sisältää sanaston, semanttiset linkitykset sekä yksinkertaisia logiikkasääntöjä. Esimerkiksi käy ontologia liittyen ruuanlaittoon ja keittokirjoihin, joka pitää sisällään esimerkiksi tarvittavat ainekset, tiedon siitä kuinka ne tulisi sekoittaa keskenään sekä eroavaisuudet hauduttamisen ja uppopaistamisen välillä. Samoin siihen sisältyy tietoa moniselitteisestä käsitteestä öljy, jota tässä tapauksessa käytetään paistamiseen - ei moottorin voiteluun.

Ontologioiden kehitys lähti alkujaan liikkeelle tarpeesta laatia jaettuja ja uudelleenkäytettäviä tietämuskantoja. Ontologioiden avulla saadaan käyttöön yhteinen kieli, jolloin kaikki osapuolet ovat samaa mieltä termeistä ja ominaisuuksista. Uudelleenkäyttö ja jakaminen eivät ole kuitenkaan ongelmattomia, sillä käyttäjät eivät välttämättä jaa ontologian suunnittelijan implementoimia näkemyksiä ja olettamuksia käsiteltävästä aiheesta. Esimerkiksi yksi ontologia voi esittää värin punainen *suhteeksi*, kun taas toinen esittää sen *arvona*. (Gruninger & Lee 2002)

Gruninger ja Lee (2002) ovat osittaneet ontologioiden käytön kolmeen eri osaan, joita ovat kommunikointi, koneellinen päättely sekä tietämyksen järjestäminen ja uudelleenkäyttö. Kommunikointia tapahtuu tietojärjestelmien kesken, ihmisten välillä sekä ihmisten ja tietojärjestelmien vaihtaessa tietoa. Koneellisessa päättelyssä ontologioita käytetään tiedon esittämiseen ja käsittelyyn sekä tieto-

järjestelmien sisäisten rakenteiden, algoritmien, syöttö- ja tulostietojen teoreettiseen ja käsitteelliseen analysointiin. Tietämyksen hallinnassa ontologioita käytetään tietovarastojen järjestelyyn ja rakenteistamiseen.

Semanttista tietoa voidaan jo nykyisellään liittää ja käsitellä useiden, pääsääntöisesti XML-pohjaisten, tekniikoiden avulla. RDF (Resource Description Framework) ja siihen liitetty schema-kieli RDFS sopivat assosiaatioiden laatimiseen tietojen välille. XML Topic Maps (XTM) on toisenlainen mekanismi taksonomian esittämiseen informaatiolle sekä tiedon luokitteluun. (Daconta ym. 2003) Sekä Topic Maps että RDF palvelevat samaa tarkoitusperää, eli torjuvat infoähkyä WWW:n resurssien metakuvauksilla. (Hyvönen 2001) Web-palvelut (Web services) tarjoavat puolestaan mekanismin, jolla järjestelmät voivat kommunikoida keskenään. Ontologiakielet (OWL, DAML+OIL ym.) tukevat RDF:ia, esimerkiksi DAML+OIL:sta johdettu OWL tarjoaa sille sanastolaajennuksen käyttöön. Edellä mainitut kielet ovat myös sinällään käytössä useissa organisaatioissa tietämuskantojen semanttisessa järjestyksenpidossa. (Daconta ym. 2003)

1.4 Yhteenveto

Eri tietoyksiköiden vertaileminen ja ymmärtäminen yhteisellä tavalla on helpompaa tai ylipäätään mahdollista, kun kohteet käyttävät jaettua kontrolloitua sanastoa. Tiedon hakuun, esittämiseen ja mallintamiseen liittyvät järjestelmät hyötyvät ontologioista. Metakuvausten ja ontologiatekniikoiden tärkeitä sovel-lusalueita ovat Hyvösen (2001) mukaan esimerkiksi:

- Informaation haku (information search/retrieval)
- Tietämyksen hallinta (knowledge management)
- Verkkokauppa (web commerce)
- Sähköinen liiketoiminta (electronic business).

Tesaurukset ja ontologiat ovat hyödyllisiä tietojen kuvailemisessa ja luokittelus-sa. Lupaava sovellusalue on esimerkiksi jo aiemmin mainittu World Wide Web, johon semanttisen sisällön lisääminen tuo paitsi helpotusta tietoa hakeville ih-miskäyttäjille, myös mahdollisuuksia koneelliseen käyttöön (erilaiset agentit, web-mönkijät ym.) Semanttinen web lunastanee lupauksensa siinä vaiheessa, kun standardoinnit etenevät ja normaalikäyttäjän ei tarvitse välittää enää mata-lan tason merkkauksesta.

LÄHDELUETTELO

- Baeza-Yates R. & Ribeiro-Neto B. 1999. Modern Information Retrieval. Addison Wesley. New York: ACM Press.
- Chandrasekaran B., Josephson J.R & Benjamins R. 1999. What are ontologies and why do we need them? IEEE Intelligent Systems and Their Applications 14(1), 20-26.
- Daconta M., Obrst L. & Smith K. 2003. The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. John Wiley & Sons.
- Defude B. 1984. Knowledge based systems versus thesaurus: an architecture problem about expert systems design. Teoksessa Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval, Cambridge, England. Swinton: British Computer Society, 267 – 280.
- Gruninger M. & Lee J. 2002. Ontology applications and design. Communications of the ACM 45(2), 39-41.
- Helsingin yliopiston kirjasto. 2006. VESA - verkkosanasto/webbtesaurus – YSA – Yleinen suomalainen asiasanasto [online], Helsinki: Helsingin yliopiston kirjasto [viitattu 22.5.2006]. Saatavissa [www-muodossa <http://vesa.lib.helsinki.fi/ysa/index.html>](http://vesa.lib.helsinki.fi/ysa/index.html).
- Hendler J. 2001. Agents and the Semantic Web. IEEE Intelligent Systems 16(2), 30-37.
- Hyvönen E. 2005. Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita? Signum 5/2005.

Hyvönen E. 2001. Semantic Web – kohti uutta merkitysten Internetiä. Esitelmä
Semantic Web Kick-Off in Finland -tilaisuudessa 2.11.2001. Helsinki:
Helsingin yliopisto ja Helsinki Institute for Information Technology
(HIIT).

Järvelin K. 1995. Tekstitiedonhaku tietokannoista: johdatus periaatteisiin ja
menetelmiin. Espoo: Suomen atk-kustannus.