TEKNILLINEN KORKEAKOULU
Tietotekniikan osasto
Informaatioverkostojen tutkinto-ohjelma

Antti Vehviläinen

Ontologiapohjainen kysymys-vastauspalvelu

Diplomityö, joka on jätetty opinnäytteenä tarkastettavaksi
diplomi-insinöörin tutkintoa varten

Espoon Otaniemessä 11.10.2006

Työn valvoja ja ohjaaja: Professori Eero Hyvönen

Teknillinen korkeakoulu
Tietotekniikan osasto

Diplomityön tiivistelmä

Tekijä:	Antti Vehviläinen
Osasto:	Tietotekniikan osasto
Pääaine:	Mediatekniikka informaatioverkostoissa
Sivuaine:	Tietoliikenneohjelmistot
Työn nimi:	Ontologiapohjainen kysymys-vastauspalvelu
Title in English:	An ontology-based help desk service
Professuurin koodi ja nimi:	AS-75 Viestintätekniikka
Työn valvoja:	Prof. Eero Hyvönen
Työn ohjaaja:	Prof. Eero Hyvönen

Tiivistelmä:

Tässä diplomityössä tutkittiin, miten semanttisen webin tekniikoita voidaan hyödyntää kysymys-vastauspalveluissa. Työssä keskityttiin vastauksien laatijan rooliin palveluissa. Aluksi selvitettiin, millaista ongelmallisuutta kysymys-vastauspalveluihin sisältyy ja miten ontologiapohjaista asiasanoitusta voidaan käyttää semanttisen yhteensopivuuden saavuttamiseksi. Lisäksi tutustuttiin siihen, miten tapauksiin perustuvan päättelyn avulla voidaan luoda ratkaisuja uusiin ongelmiin vanhojen ratkaisujen perusteella.

Työssä laadittiin teoreettinen malli siitä, miten semanttista asiasanoitusta ja ontologioita voidaan yhdistää tapauksiin perustuvan päättelyn askeliin. Työtä varten kerättiin kirjastonhoitajien käyttäjätarpeita. Näiden tarpeiden ja teoreettisen mallin perusteella kehitettiin Opas, olemassa olevaan Kysy kirjastonhoitajalta -palveluun perustuva prototyyppi ontologiapohjaisesta kysymys-vastauspalvelusta. Oppaaseen kehitettiin semanttiseen tiedoneristykseen pohjautuva asiasanojen ehdottaja, ja lisäksi vastauksen kirjoittamisen avustamiseksi eri tietolähteitä yhdistettiin Oppaaseen Yleisen suomalaisen ontologian avulla. Lopuksi Opasta testattiin tekemällä käyttäjätestejä kirjastonhoitajien kanssa.

Työssä selvisi, että asiasanaehdotukset helpottavat vastausten laatijan työtä asiasanojen valinnassa, ja että ontologioissa määriteltyjä semanttisia suhteita voidaan käyttää asiasanaehdotusten relevanssin selvittämisessä. Asiasanaehdotuksia voidaan myös käyttää samankaltaisten kysymysten ja tietoresurssien etsinnässä, mikä auttaa vastauksen laatimisessa. Työssä aukesi paljon jatkotutkimuksen aiheita semanttisen webin hyödyntämisestä kysymys-vastauspalveluissa.

Sivumäärä: 85 Avainsanat: semanttinen web, tapauksiin perustuva päättely, ontologiat, semanttinen asiasanoitus, kysymys-vastauspalvelut

Täytetään osastolla

Hyväksytty: Kirjasto:

Helsinki University of Technology
Department of Computer Science and Engineering

Abstract of Master's thesis

Author:	Antti Vehviläinen
Department:	Department of Computer Science and Engineering
Major subject:	Media Technology in Information Networks
Minor subject:	Telecommunications Software
Title:	An ontology-based help desk service
Title in Finnish:	Ontologiapohjainen kysymys-vastauspalvelu
Chair:	AS-75 Media Technology
Supervisor:	Prof. Eero Hyvönen
Instructor:	Prof. Eero Hyvönen

Abstract:

The aim of this master's thesis was to study how semantic web technologies can be utilized in help desk services. The work focused on the answerer's role in the services. The work begun by studying the characteristics of help desk services, how semantic web technologies can be applied to them, and how semantic indexing can be used to achieve semantic interoperability. The theory of Case-based Reasoning was investigated to find out how old problems can be used in solving new problems.

At first, theoretical model of combining semantic indexing and ontologies with the steps of case-based reasoning was developed. Then, librarians were interviewed to collect user requirements. A prototype of an ontology-based help desk service, Opas, was developed based on the user requirements, the theoretical model and the existing Ask a Librarian -service. A semantic information extractor was used in Opas to produce index words suggestions. To help the authoring process different information resources were integrated to Opas by using the Finnish common upper ontology. Finally, Opas was evaluated with user tests.

The main results of the work are: 1) index word suggestions are useful in choosing appropriate index words for a question-answer pair, 2) semantic relations defined in ontologies can be used in determining the likely relevance of an index word suggestion, 3) index word suggestions can be used to search for similar old questions and information resources and 4) these similar old questions and different information resources can be used to author a new answer. This thesis revealed a number of insights for future research.

Number of pages: 85 Keywords: semantic web, case-based reasoning, ontologies, semantic indexing, help-desk services

Department fills

Approved: Library code:

Esisanat

Tämä diplomityö tehtiin Teknillisessä korkeakoulussa viestintätekniikan laboratorion semanttisen laskennan tutkimusryhmässä. Haluan kiittää professori Eero Hyvöstä työni kannustavasta ohjaamisesta ja hyvistä neuvoista sekä työtovereitani inspiroivista kahvihuone- ja käytäväkeskusteluista. Erityisesti haluan kiittää Stina Westmania, joka vapaaehtoisesti tarjoutui antamaan palautetta työstäni, sekä Olli Almia, jonka toteuttama ohjelma toimi suurena osana omaa työtäni.

Espoossa 11.10.2006 Antti Vehviläinen

Sisältö

Lyhenteet ja termit

CBR	Case-based Reasoning
HKLJ	Helsingin kirjaston luokitusjärjestelmä
NLP	Natural Language Processing
Opas	Työssä rakennettavan prototyypin työnimi
OWL	Web Ontology Language
Poka	SeCo-tutkimusryhmässä kehitetty työkalu semanttiseen annotointiin
RDF	Resource Description Framework
SeCo	Semantic Computing Research Group, semanttisen laskennan tutkimusryhmä
UKK	Usein Kysyttyjä Kysymyksiä
tf-idf	term frequency - inverse document frequency
WWW	World Wide Web
XML	eXtensible Markup Language
YSA	Yleinen suomalainen asiasanasto
YSO	Yleinen suomalainen ontologia

Kuvat

Luku 1

Johdanto

Yritykset ja julkiset järjestöt hyödyntävät paljon tukikeskuspalveluita (engl. *help desk services*) ratkaistakseen asiakkaidensa ongelmia. Perinteisin esimerkki näistä palveluista ovat auttavat puhelimet, joihin asiakkaat voivat soittaa ongelmatilanteissa. World Wide Webin (WWW, web) käytön yleistymisen myötä myös tukikeskuspalveluita siirretään webiin, ja yhä useammin asiakkaat voivat ratkaista ongelmansa itse web-palvelun kautta ottamatta suoraan yhteyttä tukihenkilöön [FHLL00].

Esimerkki yksinkertaisista ja helposti toteutettavista webissä olevista tukikeskuspalveluista ovat niin sanotut UKK-listaukset (Usein Kysytyt Kysymykset), joihin on listattu tietystä aihealueesta usein esitettyjä kysymyksiä sekä vastauksia niihin. Usein käyttäjät voivat myös lähettää tietyn alan asiantuntijoille kysymyksiä, ja vastauksia voi selata palvelun arkistossa. Tämä raportti käsittelee näitä kysymys-vastauspalveluita.

Sen lisäksi, että käyttäjät arvostavat sitä, että voivat itse etsiä vastauksen ongelmaansa web-palvelun kautta, ovat tukikeskuspalvelut hyödyllisiä myös palvelun tukihenkilön näkökulmasta. Ensiksi, jos asiakas saa ratkaistua ongelman itse, säästyy tukihenkilön aikaa ja resursseja. Toiseksi, tukihenkilö voi itse hyödyntää palvelun tietokantaa vastatessaan kysymyksiin.

Semanttinen web [BLHL01] on Tim Berners-Leen visio siitä, minkälainen tulevaisuuden Webin pitäisi olla. Semanttisen webin tavoitteena on se, että ihmisten lisäksi myös koneet ymmärtäisivät WWW-sivujen sisältöjä, eivätkä pitäisi niitä vain merkkijonojoukkona. Tämän voi mieltää siten, että jos nykyinen web on suuri web-sivujen kirjasto, niin tulevaisuuden semanttinen web olisi jättimäinen web-sivujen tietokanta. Tämä vaatii muun muassa sen, että web-sivuihin liitetään koneellisesti ymmärrettävää metatietoa, annotaatioita.

Yksi semanttisen webin kehityksen suurimmista jarruista on se, että palvelut tarvitsevat sivuihin liitettyä metatietoa, mutta koska palveluita ei ole paljon, järjestöt, yritykset ja ihmiset eivät välttämättä näe syytä luoda tätä metatietoa, kuten myös Dill ym. [DEG+03] toteavat. Kehitystä jarruttaa heidän mukaan myös se, että suuri osa tämänhetkisestä webin sisällöstä, metatieto mukaanlukien, on tehty luonnollista kieltä käyttäen. Jotta semanttinen web voisi lyödä itsensä lopullisesti läpi, olisi semanttista metatietoa oltava saatavilla sekä uusia että vanhoja web-dokumentteja varten [RH05].

Kirjastoala on perinteisesti ollut hyvin metatietorikasta. Kirjastojen aineistot luokitellaan tarkasti, ja alan ammattilaiset saavat koulutuksen metatiedon luomista varten. Kirjastoala onkin erityisen kiinnostava metatietoon nojautuvan semanttisen webin näkökulmasta. Metatietorikkaita aineistoja on helppo muuntaa semanttisen webin sovellusten tarpeisiin, ja eri aineistoja voidaan yhdistää kirjastoalalla käytettävien luokittelujärjestelmien perusteella.

Kirjastoalalla on myös runsaasti kysymys-vastauspalveluita, joiden avulla käyttäjät voivat kysyä kirjastonhoitajilta kysymyksiä menemättä itse kirjastoon. Tällaisissa kysymys-vastauspalveluissa on tunnistettavissa joitakin samankaltaisia piirteitä ja ongelmakohtia:

- Olemassaolevan tiedon hyödyntäminen. Koska käyttäjien ongelmat toistuvat usein, on hyvä ensin tarkastaa, onko kysymykseen jo vastattu aikaisemmin. Vaikka täysin samanlaista kysymystä ei olisikaan esitetty, voisiko samankaltaisista kysymyksistä olla apua vastaustyössä?
- Ulkoisten tietolähteiden hyödyntäminen. Vastaamistyössä käytetään usein monenlaisia tietolähteitä, kuten kirjastojen käsikirjastoja ja tietokantoja. Miten näitä tietolähteitä voitaisiin yhdistää palveluun ja siten helpottaa vastaamistyötä?
- Asiasanoitus. Miten kysymys-vastauspareja varten kannattaa valita asiasanat aineiston indeksointia varten?

Kysymys-vastauspalveluissa voidaan nähdä kaksi pääkäyttäjäryhmää. Toinen on palvelua käyttävät, kysymyksiä esittävät ja arkistoa selaavat loppukäyttäjät, toinen on palvelua ylläpitävät, kysymyksiin vastaavat tukihenkilöt. Tässä työssä keskitytään tukihenkilön näkökulmaan palvelussa. Yhtenä tavoitteena on tutkia, miten semanttisen webin tekniikoiden avulla voidaan avustaa kysymyksiin vastaamisessa. Toisena tavoitteena on tutkia yleisemmällä tasolla, miten semanttisen webin tekniikoita voitaisiin hyödyntää erilaisissa kysymys-vastauspalveluissa.

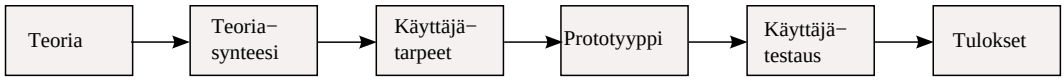
Tämä raportti pyrkii vastaamaan seuraaviin tutkimuskysymyksiin:

1. *Miten semanttisen webin tekniikoita voidaan hyödyntää helpottamaan vastaamistyötä kysymys-vastauspalveluissa?*
2. *Millaisia uusia mahdollisuuksia semanttisen webin tekniikat avaavat kysymys-vastauspalveluiden toteuttamisessa?*

Työssä käytetään konstruktivistista tutkimusotetta. Olemassaolevaa teoriaa yhdistämällä ja analysoimalla luodaan teoreettinen synteesi siitä, miten kysymys-vastauspalveluita kannattaisi rakentaa semanttisen webin tekniikoita hyödyntäen. Luodun teoreettisen mallin testaamiseksi laaditaan käyttäjätarpeiden perusteella konstruktio, prototyyppi, jonka avulla mallia testatataan empiirisesti olemassaolevalla datalla ja käyttäjillä.

Tämän raportin teoriaosa rakentuu siten, että aluksi kappaleessa 2 pureudutaan ongelma-alueeseen tutustumalla siihen, mitä kysymys-vastauspalvelut ylipäänsä ovat , ja siihen miten niitä voidaan toteuttaa. Kappaleessa myös tutustutaan semanttiseen webiin, ja miten ontologioita voidaan käyttää kysymys-vastauspalveluissa. Tämän jälkeen kysymys-vastauspalveluiden piirteitä tarkastellaan yksityiskohtaisemmin kysymyksiin vastaajan näkökulmasta: kappaleessa 3 käsitellään kysymysten ja vastausten asiasanoitusta, ja kappaleessa 4 kerrotaan, miten tapauksiin perustuvaa päättelyä (engl. *Case-based Reasoning*) voidaan käyttää samankaltaisten kysymysten etsimisessä. Teoriaosuus vedetään yhteen kappaleessa 5, jossa esitetään malli ontologiapohjaisesta semanttista asiasanoitusta käyttävästä kysymys-vastauspalvelusta.

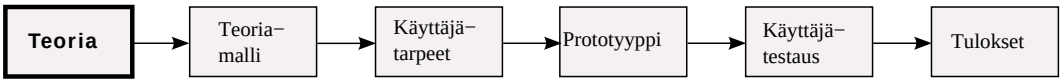
Raportin empiirinen osa alkaa kappaleesta 6, jossa esitellään työtä varten tehty käyttäjätarpeiden määrittely. Tähän määrittelyyn ja teoriasynteesiin perustuva prototyyppi esitellään kappaleessa 7. Prototyypin avulla tehdyt käyttäjätestit esitellään kappaleessa 8. Käyttäjätestien tulokset esitellään kappaleessa 9, jossa myös analysoidaan toteutettua prototyyppiä ja vastataan tutkimuskysymyksiin. Lopuksi esitetään jatkotutkimuksen aiheita ja vedetään raportti yhteen johtopäätösten muodossa kappaleessa 10. Raportin kulkua on havainnollistettu kuvassa 1.1.



Kuva 1.1: Työn rakenteen kuvaus

Luku 2

Kysymys-vastauspalvelut



Tämän kappaleen tarkoituksena on tutustuttaa lukija erilaisiin kysymys-vastauspalveluihin. Lisäksi tutustutaan, miten tietämystä esitetään semanttisessa webissä ontologioiden avulla, ja miten semanttisen webin tekniikoita voidaan käyttää kysymys-vastauspalveluissa.

2.1 Kysymyksiin ja vastauksiin perustuvat järjestelmät

Erilaiset kysymyksiin ja vastauksiin perustuvat järjestelmät voidaan karkeasti jakaa kahteen osaan. *Kysymys-vastauspalveluilla* tarkoitetaan palveluita, joiden aineisto koostuu luonnollisella kielellä esitetyistä kysymyksistä ja vastauksista. Esimerkkejä tällaisista palveluista ovat yritysten tukikeskuspalvelut tai UKK-listaukset, joissa on listattuna tiettyyn aihealueeseen liittyviä usein toistuvia kysymyksiä. *Kysymyksiinvastaamisjärjestelmät* (engl. *Question-Answering System*) taas ovat tiedonhakujärjestelmiä, jotka muodostavat luonnollisella kielellä esitettyyn kysymykseen vastauksen sen sijaan, että vain etsisivät kysymykseen liittyviä dokumentteja [Voo01].

Kysymys-vastauspalvelut ovat oikeastaan kysymyksiinvastaamisjärjestelmän erikoistapaus. Kun kysymys-vastauspalvelun aineistoon tehdään hakuja, voi

järjestelmä "muodostaa" vastauksen etsimällä kysymys-vastausaineistosta vastauksia esitettyyn kysymykseen. Erona kysymyksiinvastaamisjärjestelmiin on kuitenkin se, että kysymys-vastauspalveluissa aineisto on jo valmiina kysymys-vastausmuodossa, kun taas kysymyksiinvastaamisjärjestelmissä vastaus muodostetaan dynaamisesti eri lähteistä, jotka eivät välttämättä ole valmiina kysymys-vastausmuodossa. Tässä raportissa keskitytään kysymys-vastauspalveluihin.

2.1.1 Kysymyksiin vastaamisjärjestelmät

Erilaisia kysymyksiinvastaamisjärjestelmiä on tutkittu runsaasti. FAQ FINDER [BHK+97] on järjestelmä, jossa luonnollisella kielellä esitettyyn kysymykseen etsitään vastaus monista UKK-listauksista. Järjestelmä käyttää yleisesti käytetyn tf-idf-menetelmän [SM86] (engl. *term frequency - inverse document frequency*) lisäksi WordNet-sanastoa [Mil95] oikean vastauksen etsimiseen. Yhtenä tämän järjestelmän rakentamisen yhteydessä tehdyn tutkimustyön merkittävistä johtopäätöksistä mainittakoon se, että siinä missä perinteiseen tf-idf:n perustuvilla menetelmillä saadaan tyydyttäviä tuloksia, tuo WordNet-sanaston käyttäminen selkeitä etuja vastausten hakemiseen.

Webin yleistyttyä vuosituhannen vaihteen läheisyydessä tutkimuksellinen mielenkiinto on alkanut siirtyä siihen, miten kysymyksiin voidaan vastata erilaisten web-lähteiden avulla sen sijaan, että vastausjoukkona olisi ennalta rajattu joukko dokumentteja, kuten UKK-listauksia. Esimerkiksi MULDER [KEW01] on järjestelmä, jossa FAQ FINDER:ssä käytettyjä klassisia tiedonhaun menetelmiä sovelletaan vastauksen etsimiseen webistä. NSIR [RFQ+05] taas on samankaltainen järjestelmä, jossa hakumenetelmiä on kehitetty edelleen todennäköisyyslaskennan teoriaa hyödyntäen.

2.1.2 Kysymys-vastauspalvelut

Web-pohjaisia kysymys-vastauspalveluita käytetään usein yritysten ja yhteisöjen tukikeskuspalveluina [FHLL00]. Esimerkiksi Kirjastot.fi-toimitus listaa verkkosivuillaan[1] lukuisia suomalaisia kysymys-vastauspalveluita.

Marom ja Zukerman [MZ05] ovat kehittäneet järjestelmän, jossa asiakkaan sähköpostitse yritykselle lähettämään tukipyyntöön etsitään olemassa olevista tukipyynnöistä samankaltaisia sähköposteja. Järjestelmä myös päättelee, on-

[1] http://www.kirjastot.fi/tiedonhaku/kysypalveluita

ko löydetty tukipyyntö riittävän tarkka, jolloin vanha vastaus tukipyyntöön lähetetään suoraan asiakkaalle. Muussa tapauksessa ylläpitäjä muokkaa vanhaa tukipyyntöä ennen asiakkaalle lähettämistä. Kysymysten ja ongelmien ei välttämättä tarvitse olla sähköpostitse esitettyjä. Feng ym. [FSKH06] ovat kehittäneet yliopistomaailmaan keskusteluohjelman, "botin", joka etsii ratkaisuja opiskelijoiden ongelmiin keskustelupalstan viestiketjujen ja kurssisivujen perusteella.

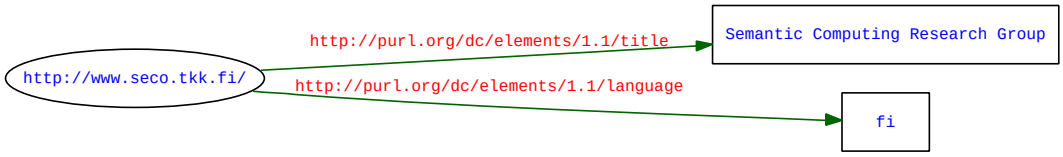
Vaikka edellä esitellyt kysymys-vastauspalvelut eroavatkin toisistaan huomattavasti, on niiden perusperiaate kuitenkin samankaltainen: käyttäjällä on jokin ongelma, kysymys, ja tähän ongelmaan pyritään löytämään ratkaisu automaattisesti tai edes puoliautomaattisesti. Ajatuksena taustalla on usein, että samankaltainen ongelma on esiintynyt joskus ennenkin. Käyttöliittymä ongelman esittämiseen ja tietovarasto, josta vastauksia etsitään, muuttuvat, mutta perusperiaate on sama. Tavoitteena on usein ajansäästö ja parempi palvelu asiakkaille. Voidaan myös ajatella, että yritysten ja yhteisöjen käyttämissä kysymys-vastauspalveluissa on palattu juurille FAQ FINDER:n aikaan. Erona on se, että vastauksia etsitään monimutkaisemmista tietolähteistä kuin UKK-listauksista, ja että nykyteknologioilla voidaan toteuttaa monimutkaisempia kysymiskäyttöliittymiä kuin 90-luvulla. Vastauksia ei myöskään pyritä etsimään koko webistä MULDER:n ja NSIR:n tapaan.

Monimutkaisissa ongelma-alueissa käyttäjien voi olla myös vaikea ilmaista ongelmiaan. Yksi olennainen osa-alue kysymys-vastauspalveluissa onkin se, miten käyttäjää tuetaan ongelmankuvausprosessissa. Delisle ja Moulin [DM02] esittävät, että järjestelmällä pitää olla tietämystä käyttäjästä, kyky tunnistaa käyttäjän ongelmia ja kyky auttaa käyttäjää ongelman kuvailussa. Nämä ominaisuudet ovat sitä tärkeämpiä, mitä tarkempaan ja teknisempään ongelmaan haetaan ratkaisua.

2.2 Semanttinen web

Nykyisessä webissä ongelmana on se, että metatieto on tehty luonnollisella kielellä, ja että yhteisiä, laajalti käytössä olevia metatietostandardeja ei ole. Semanttinen web voidaan nähdä nykyisen webin päällä olevana metatietokerroksena, ja se yrittää ratkaista semanttisen yhteensopimattomuuden ongelman. Tämän metatietokerroksen olennainen ominaisuus on *koneymmärrettävyys*. Ajatuksena on se, että erilaiset koneet pystyvät hyödyntämään metatietoa älykkäissä sovelluksissa [Av04].

Perinteisen webin metatietoratkaisut pohjautuvat usein XML-sisällonkuvauskieleen[2], kun taas semanttisessa webissä tässä roolissa toimii Resource Description Framework[3] (RDF). RDF-kielen avulla voidaan esittää niin sanottuja lauseita, väittämiä (engl. *statement*), jotka koostuvat luonnollisen kielenkäytön lauseiden tapaisesti subjektista, predikaatista ja objektista. Tästä kolmikosta käytetään myös nimitystä *tripletti*.



Kuva 2.1: Esimerkki RDF-graafista

RDF:n lauseita käytetään erityisesti kuvailemaan webissä olevia resursseja. Siinä missä XML-dokumentti muodostaa puun, muodostaa joukko RDF-lauseita verkon. Kuvassa 2.1 on havainnollistettu kahta lausetta, jotka muodostavat yksinkertaisen verkon, *graafin*. Kuvassa on kuvailtu SeCo-tutkimusryhmän www-sivuja kertomalla sivuston kieli ja otsikko. RDF-dokumenttien graafimuoto tukee myös ajatusta siitä, miten semanttisen webin metatietokerros voisi koostua valtavasta verkostosta toisiinsa viittaavia RDF-graafeja.

2.2.1 Ontologiat

RDF itsessään on rajoittunut kieli resurssien kuvailuun. Semanttisen webin tarpeisiin tarvitaan semanttisesti rikkaampia käsiterakenteita, *ontologioita*. Filosofiassa sanalla ontologia tarkoitetaan oppia olevaisesta, mutta tietojenkäsittelytieteessä sanalla on eri merkitys. Heflinin [Hef] mukaan ontologia määrittelee termit, joilla jotain tiettyä tietämyksen aluetta kuvaillaan ja esitetään. Termien lisäksi ontologioissa on esitetty formaalisti määriteltyjä käsitteitä, niiden välisiä suhteita ja erilaisia arvorajoituksia. Esimerkiksi jossain ontologiassa voisi olla määritelty, että *isä* on *vanhemman* tarkempi käsite, ja että *lapsella* voi olla vain yksi biologinen isä.

Yksi ontologioiden päätehtävistä on tietämyksen esittäminen. Randall ym. [RDS93] jakavat edelleen tietämyksen esittämisen viiteen rooliin:

[2]http://www.w3.org/XML/
[3]http://www.w3.org/RDF/

Tietämyksen esittäminen todellisuuden kuvaajana Tässä tietämyksen esittämisen roolina on kuvata todellisuuden ilmiöitä mahdollisimman tarkasti.

Tietämyksen esittäminen ontologisena sitoumuksena Edellisessä kohdassa mainittuja todellisuuden kuvauksia on lukematon määrä. Tässä roolissa tietämyksen esittämisen tehtävänä on tehdä valinta siitä, minkälainen näkökulma todellisuuteen otetaan.

Tietämyksen esittäminen osittaisena teoriana älykkäästä päättelystä Tässä tietämyksen esittämisen roolina on toimia älykkään päättelyn perustana. Tietämyksen esitys nähdään silmälläpitäen sitä, miten ihmiset tekevät päättelyä ja järkeilyä.

Tietämyksen esittäminen tehokkaan laskennan välineenä Tässä teknisessä roolissa tietämyksen esittämisen tehtävänä on toimia välineenä tehokkaalle laskennalle. Roolissa olennaista on laskennallinen vaativuus: miten esityksistä tehdään päätelmiä ja laskelmia tehokkasti?

Tietämyksen esittäminen ihmisen ilmaisun välineenä Viimeisenä tietämyksen esittämisen roolina on toimia välineenä sille, miten ihmiset ilmaisevat itseään ja kommunikoivat toistensa kanssa.

Ontologiat täyttävät nämä tietämyksen esittämisen roolit semanttisessa webissä. Olennaiseksi nousee se, miten ontologiat toimivat eräänlaisena rajapintana ihmisen ja koneen välillä – ihmiset voivat intuitiivisesti kuvata todellisuutta ontologioiden avulla ja tämän kuvauksen perusteella voidaan tehdä koneellista päättelyä ja tehokasta laskentaa. Huomattavaa on myös se, miten tietyn ontologian käyttö tietystä aihepiiristä on myös sitoumus tietyn termistön käytöstä.

2.2.2 Ontologiakieli OWL

Ontologioiden määrittelyä varten tarvitaan jokin kieli. Web Ontology Language [MvH] (OWL) on World Wide Web Consortiumin (W3C) suosittelema kieli webissä käytettävien ontologioiden määrittelemistä varten. Kielestä on määritelty kolme eri versiota, joiden ominaisuudet vaihtelevat ilmaisuvoiman ja laskennallisen vaativuuden suhteen. *OWL Lite* on versioista kevein, ja se on tarkoitettu esimerkiksi yksinkertaisten sanastojen määrittelyyn, silloin kun ei tarvita koneellista päättelyä. *OWL DL* on ilmaisuvoimaltaan mahdollisimman runsas kuitenkin ollen laskennallisesti ratkeava, eli sen avulla tehdyistä ontologioista voidaan aina tehdä päätelmiä äärellisessä ajassa. *OWL Full* taas on

ilmaisuvoimaltaan runsain ontologiakieli, mutta sillä ei ole *OWL DL*:n kaltaista takuuta laskennallisesta ratkeavuudesta.

2.2.3 Olemassa olevien aineistojen ontologisointi

Ei ole realistista, että semanttisen webin tarpeita varten erilaiset aineistot ja sanastot luotaisiin uudelleen tyhjästä. Tästä syystä olemassa olevat aineistot on muunnettava semanttisen webin sovellusten ymmärtämään RDF/OWL-muotoon. Tästä prosessista käytetään termiä *ontologisointi*.

Van Assem ym. [vAMS+04] ovat kehittäneet menetelmän erilaisten sanastojen muuntamiseen RDF/OWL-muotoon. Siinä sanastojen muuntaminen on jaettu neljään askeleeseen: valmistelu, syntaktinen muunnos, semanttinen muunnos ja suhteutus johonkin standardiin:

1. **Valmistelu.** Ennen sanaston muuntamista RDF/OWL-muotoon on syytä ottaa huomioon alkuperäisen sanaston suhde erilaisiin standardeihin ja monikielisyyteen.

2. **Syntaktinen muunnos.** Tällä askeleella tarkoitetaan prosessia, jolla sanasto muunnetaan lähdemuodosta (esimerkiksi XML- tai relaatiotietokantaesitys) RDF/OWL-muotoon. Muunnoksessa käytetään vain yksinkertaisia RDF/OWL-rakenteita. Askeleessa on muistettava muun muassa välttää tulkintojen tekemistä ja toisteisen tiedon lisäämistä. Olennaista siis on, että olemassa oleva esitys muunnetaan sellaisenaan RDF/OWL-muotoon lisäämättä kuitenkaan uusia semanttisia suhteita tai poistamatta vanhoja.

3. **Semanttinen muunnos.** Tässä askeleessa edellä tehtyä "raakaa" muunnosta rikastetaan semanttisilla suhteilla edellistä askelta monimutkaisemmilla RDF/OWL-rakenteilla. Tässä askeleessa voidaan tehdä myös tulkintoja, jota alkuperäisessä aineistossa ei ollut.

4. **Muunnoksen suhteutus johonkin standardiin.** Kun muunnos on tehty, tuloksena saatu ontologia suhteutetaan johonkin olemassa olevaan standardiin. Esimerkiksi Simple Knowledge Organisation System[4] (SKOS) on W3C:n suositusskeema erilaisia sanastoja varten. Suhteuttamisessa tehdään esimerkiksi "muunnosontologia", joka kertoo, miten oman ontologian käsitteet ja suhteet suhtautuvat standardiontologiaan. Näin muut, standardiontologiaa tukevat mutta omaa ontologiaa

[4]http://www.w3.org/2004/02/skos/

ymmärtämättömät sovellukset osaavat käyttää tehtyä ontologiamuotoista sanastoa.

Vaikka Van Assem ym. ovatkin valmistelleet edellä mainitut ohjenuorat erityisesti sanastojen muuntamista varten, voitaneen niitä soveltaa myös muiden aineistojen muuntamiseen.

2.3 Kysymys-vastauspalvelut semanttisessa webissä

Semanttisen webin teknologioita ollaan hyödynnetty erilaisissa kysymyksiinvastaamisjärjestelmissä. PowerAqua [LMU06] on ontologiapohjainen järjestelmä, joka muodostaa luonnollisella kielellä esitettyyn kysymykseen vastauksen erilaisten semanttisten resurssien perustella. Lopez ym. painottavat, että ontologiat ovat olennainen osa semanttista webiä: ne mahdollistavat semanttisen yhteentoimivuuden, automaattisen informaation käsittelyn sekä sen, että erilaisia heterogeenisiä aineistoja voidaan yhdistää.

Kyselyn laajentaminen [XC96] (engl. *query expansion*) on tekniikka, jolla käyttäjän esittämää kyselyä laajennetaan esimerkiksi synonyymeillä. Tämä tehdään siksi, että usein käyttäjän sanasto eroaa termistöstä, jolla järjestelmässä oleva aineisto on kuvattu. Bilotti [Bil04] on tutkinut kyselyn laajentamista kysymys-vastausjärjestelmissä ja esittää, että tekniikka parantaa vastausten saantia, mutta kyselyä ei kannata laajentaa liikaa, koska tällöin saadaan paljon myös epäolennaisia vastauksia.

Kyselyjä on laajennettu erilaisissa järjestelmissä jo ennen semanttisen webin tuloa, mutta ontologiat mahdollistavat älykkään kyselyn laajentamisen. Esimerkiksi PowerAqua hyödyntää ontologioita muun muassa tunnistaessaan mitä tietyllä sanalla tarkoitetaan tietyssä yhteydessä. Esimerkiksi englanninkielisessä kysymyksessä *What is the capital of Spain?* (Mikä on Espanjan pääkaupunki?) monimerkityksellisen sanan *capital* (mm. pääkaupunki, pääoma, iso kirjain) merkitys päätellään siitä, että sanan *Spain* tiedetään ontologian perusteella olevan maa, ja käsitteillä *maa* ja *pääkaupunki* on käytetyssä ontologiassa semanttinen suhde.

Suuri osa PowerAquan toiminnallisuutta on myös se, miten käyttäjän käsitteistö peilataan sovelluksen ontologioiden käsitteistölle. Lopez ym. [LMU06] esittävät, että käyttäjän ei pitäisi joutua tekemään valintoja moniselitteisten käsitteiden välillä, koska käyttäjä ei osaa erotella, mitä

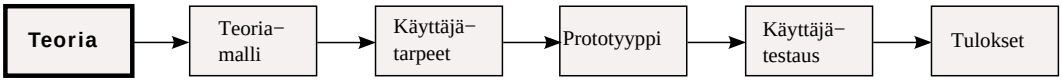
käsite tarkoittaa eri ontologioissa. Tätä ongelmaa voi kuitenkin lähestyä siten, että sovellus käyttää käsitteiden tunnistamiseen jotain yleiskieleen perustuvaa ontologiaa, ja sovelluksen käyttämät tarkemmat ontologiat on sidottu tähän yleisontologiaan. Tämä ei täysin poista moniselitteisten käsitteiden erottelua käyttäjältä, mutta merkityksen valinta yleisestä ontologiasta on helpompaa kuin valinta monen tarkemman ontologian välillä.

Erilaisissa kysymys-vastauspalveluissa voidaan tunnistaa ainakin kaksi ongelmakohtaa: 1) miten auttaa käyttäjää kysymyksen esittämisessä ja 2) miten etsiä vastauksia tehokkaasti kysymyksen perusteella. Semanttinen web tuo ratkaisuja molempiin näistä kohdista. Käyttäjän on hyvä antaa esittää kysymys luonnollisellä kielellä, jonka jälkeen ontologioiden ja niissä esitettyjen semanttisen suhteiden perusteella voidaan älykkäästi päätellä, mitä käyttäjä oikeastaan tarkoitti. Tämän jälkeen, kun erilaiset vastausarkistot ovat sidoksissa käytettyihin ontologioihin, vastauksen etsiminen helpottuu.

On syytä muistaa, että semanttinen web ja ontologiat eivät sinänsä mahdollista mitään sellaista, mikä ei aiemmin olisi jo ollut mahdollista; koneellista päättelyä ja erilaisten aineistojen integrointia on voitu tehdä jo aiemminkin. Semanttinen web ja sen ontologiat tekevät siitä vain helpompaa.

Luku 3

Asiasanoitus



Edellisessä kappaleessa tutustuttiin kysymys-vastauspalveluihin ja siihen, miten niissä voidaan hyödyntää semanttisen webin tekniikoita. Tässä kappaleessa käsitellään aineiston indeksointia kysymys-vastauspalveluissa: miten kysymys-vastausaineisto kannattaa asiasanoittaa, ja mitä ontologiat tuovat asiasanoitukseen.

3.1 Terminologiaa

Tietojärjestelmien sisältö voidaan jakaa karkeasti kahteen osaan: varsinaisen tiedon sisältävät dokumentit sekä metatieto näistä dokumenteista, kuten esimerkiksi luontiaika, koko ja niin edelleen. Yksi metatietotyypeistä ovat *asiasanat*, joilla tarkoitetaan termejä ja tunnisteita, jotka kuvaavat dokumenttien sisältöä [SM86].

Semanttisen webin yhteydessä puhutaan usein myös *annotaatioista*. Kahan ja Koivunen [KK01] määrittelevät annotaatiot metatiedoksi, jotka liittävät kommentteja tai huomautuksia johonkin dokumenttiin. Tästä metatiedon liittämisprosessista käytetään termiä *annotointi*. Asiasanoitus voidaan siis nähdä oikeastaan annotoinnin erikoistapauksena. Asiasanoituksen tarkoituksena on kuvailun lisäksi myös luokitella aineistoa.

3.2 Kontrolloitu ja vapaa asiasanoitus

Yksi avoin kysymys asiasanoituksessa on se, kannattaako käyttää kontrolloitua sanastoa vai sallitaanko käytettävien asiasanojen vapaa valinta. Useat asiantuntijat [SM86] ovat sitä mieltä, että luonnollisen kielen luonteen takia vapaa asiasanoitus altistaa väärinkäsityksille ja virheille. Kontrolloidun sanaston käyttö vähentää näin esimerkiksi synonyymeistä aiheutuvia virheitä, kun käsitteistä käytetään vain tiettyjä, ennalta sovittuja termejä. Lisäksi kirjoitusvirheet voidaan eliminoida.

Joidenkin tutkimusten mukaan kontrolloidun sanaston käyttäminen ei tehosta hakuja verrattuna vapaan asiasanoituksen käyttöön [Mud98, HBLH94]. Srinivasan [Sri96] kuitenkin väittää, että MEDLINE-palvelussa[1] kontrolloidun sanaston käyttö tuo etuja vapaaseen asiasanoitukseen nähden. Srinivasanin mukaan aineistohaku on tehokkainta, kun haussa käytetään yhdessä kontrolloitua sanastoa sekä vapaita hakutermejä. Vaikuttaakin siltä, että sovellusten kannattaa yhdistää vapaan asiasanoituksen ja kontrolloidun sanaston käyttämisen ominaisuuksia.

3.3 Automaattinen ja puoliautomaattinen asiasanoitus

Perinteisesti dokumentteja on asiasanoitettu manuaalisesti. Tällöin asiasanoittaja valitsee käsin haluamansa sanat vapaasti tai jostain sanastosta. Manuaalista asiasanoitusta ei pidä nähdä kuitenkaan täysin vapaana, vaan siinäkin on syytä noudattaa tiettyjä sääntöjä esimerkiksi valittujen asiasanojen määrän suhteen [SM86]. Manuaalinen asiasanoitus on kuitenkin vaikeaa, kallista ja aikaa vievää [CDPW02]. Etenkin semanttisessa webissä, jossa erilaisten annotaatioiden ja asiasanoitusten luominen on koko teknologian yleistymisen edellytys, tarvitaan automaattisia menetelmiä asiasanoitusten ja annotaatioiden luomiseen.

3.3.1 Automaattinen asiasanoitus

Automaattisella asiasanoittamisella tarkoitetaan sitä, että asiasanoittamisen hoitaa ihmisen puolesta tietokoneohjelma [Arm00]. Siinä tekstistä et-

[1] `http://medline.cos.com/`

sitään käsitteitä ja lauseenjäseniä luonnollisen kielen prosessointimenetelmillä (engl. *Natural Language Processing, NLP*). Löydetyt käsitteet liitetään tämän jälkeen automaattisesti dokumentteihin, ja ne muodostavat dokumentin *yhteenvedon*. Automaattisessa asiasanoituksessa pitää myös huomioida se, mitä varten asiasanoitus ja yhteenveto tehdään. Esimerkiksi Hahn ja Mani [HM00] erottelevat dokumenttien yhteenvedon kolme roolia: yhteenveto tiedonhakua varten, informatiivinen yhteenveto ja dokumenttia arvioiva yhteenveto.

Automaattinen asiasanoittaminen, yhteenvetojen laatiminen ja luonnollisen kielen prosessointi sisältävät monimutkaisia menetelmiä, eikä niiden teoriaa käsitellä tässä työssä laajemmin.

3.3.2 Puoliautomaattinen asiasanoitus

Automaattisessa asiasanoituksessa on yksi suuri haaste: kuinka varmistaa laaditun asiasanoituksen laatu, eli vastaako asiasanoitus todella asiasanoitetun dokumentin sisältöä? Yksi ratkaisu tähän on käyttää *puoliautomaattista asiasanoitusta* [HLB99, EMSS00, RH05], jolla tarkoitetaan sitä, että asiasanoitukseen osallistuu sekä ihminen että kone. Erilaisia puoliautomaattisia asiasanoitusjärjestelmiä on lukuisia, mutta yhteistä näille on se, että asiasanojen etsimiseen käytetään automaattisen asiasanoituksen menetelmiä, mutta varsinaisen päätöksen asiasanan käyttämisestä tekee ihminen. Tällöin puoliautomaattisen asiasanoituksen tehtävänä on päätellä dokumentin tekstin perusteella mahdollisia asiasanoja, ja ehdottaa niitä käyttäjälle.

Puoliautomaattisessa asiasanoituksessa saavutetaan se etu, että asiasanat ovat vahvistettuja ja varmasti relevantteja dokumentin kannalta, olettaen että asiasanoitusta tekevä henkilö tekee työnsä ammattitaitoisesti. Toisaalta, jos ajatellaan semanttista webiä laajassa mittakaavassa ja annotointia yleisesti, on annotointiprosessi syytä automatisoida mahdollisimman pitkälle – metatiedon lisäämiseen käytetty aika on kuitenkin aina pois varsinaisen tiedon tuottamiselta. Ongelma on pienempi sovelluksissa, joissa asiasanoitettavan aineiston määrä ei ole suuri, ja joissa ei käytetä skeemapohjaista annotaatiota – muutaman asiasanan lisääminen on helpompaa kuin monimutkaisen annotaatioskeeman mukaisen metatiedon lisääminen.

3.3.3 Asiasanaehdotusten järjestäminen

Puoliautomaattisessa asiasanoituksessa on se ongelma, että jos dokumentti on pitkä, voi ehdotettavia asiasanoja olla käytettävästä tiedoneristys-

menetelmästä riippuen olla paljonkin. Lisäksi kaikki ehdotetut sanat eivät välttämättä ole relevantteja dokumentin sisältöön nähden. Tarvitaan siis jonkinlaisia menetelmiä asiasanojen suodattamiseen ja järjestämiseen.

Gazendamin ym. [LGB06] tekemässä televisio-ohjelmien annotointiin liittyvässä tutkimuksessa on esitelty tapa järjestää järjestelmän ehdottamia asiasanoja. Tutkimuksessa esitetään, että jos luonnollisen kielen prosessointimenetelmillä löydetään tekstistä termejä, jotka liittyvät semanttisesti toisiinsa, ovat ne merkityksellisempiä kuin irralliset termit. Tällaista toisiinsa liittyvien termien joukkoa kutsutaan *semanttiseksi klikiksi*. Esimerkiksi käsitteet *lääkäri*, *lääke* ja *apteekki* muodostavat semanttisen klikin. Termien liittyminen toisiinsa katsottiin yleisestä audiovisuaalisen alan tesauruksesta. Tämän relevanssipainotuksen lisäksi termin frekvenssi lisää sen painoarvoa suosituksessa.

Semanttisten klikkien lisäksi asiasanasuosituksia voi järjestää edellisten asiasanoitusten pohjalta. Jos esimerkiksi termin *kissa* kanssa on käytetty useassa dokumentissa termiä *eläintenhoito* asiasanana, voidaan päätellä, että nämä termit liittyvät jotenkin toisiinsa. Holi ym. [HHL06] ovat kehittäneet menetelmän, jossa semanttiseen hakuun on yhdistetty termien suhteellisen relevanssin määritys. Siinä tf-idf-menetelmää [SM86] on laajennettu ontologiapohjaiseksi, millä saavutetaan se, että synonyymit voidaan tunnistaa saman käsitteen käytöksi. Tätä menetelmää voidaan soveltaa siten, että käänteinen termifrekvenssi lasketaan sen mukaan, kuinka monta kertaa asiasanaa on käytetty asiasanana muissa dokumenteissa. Tällöin useasti käytetyt asiasanat painuvat listalla alemmas ja vähän käytettyjä nostetaan esiin. Tämä parantanee aineiston luokittelua, kun aineisto on asiasanoitettu mahdollisimman monipuolisesti. Toisaalta, jos esimerkiksi asiasanaa *runous* on käytetty usein, niin se voi tarkoittaa sitä, että useat dokumentit käsittelevät runoutta, eikä se sinänsä ole huono asiasana. Tästä huolimatta on houkutteleva ajatus nostaa vähän käytettyjä asiasanoja listalla ylemmäs.

Itse asiassa tässä kappaleessa esitellyt asiasanojen järjestystavat ovat molemmat menetelmiä eräänlaisen semanttisen läheisyyden päättelemiseen. Toisessa etäisyys päätellään implisiittisesti valittujen asiasanojen perusteella, toisessa taas eksplisiittisillä, ontologiassa ilmaistuilla suhteilla.

3.4 Asiasanoitus tiedonhaun näkökulmasta

Yksi syy asiasanoitukselle on se, että asiasanojen perusteella aineisto voidaan luokitella hierarkioihin ja niiden perusteella voidaan tehdä hakuja aineistoon.

Luokittelukin voidaan oikeastaan nähdä hakemista tukevana toimintona, sillä sen perusteella voidaan tehdä alustavia hakuvalintoja ja luokittelun perusteella aineistoa voidaan selata. Tiedonhaun tukeminen on siis pääsyy asiasanoitukselle. Mitä asiasanojen valinnassa pitää sitten ottaa huomioon, jotta se tukee tätä hakemistoimintoa parhaiten?

Asiasanojen valinnassa on kaksi ulottuvuutta: *perusteellisuus* (engl. *exhaustivity*) ja *tarkkuus* (engl. *specificity*). Perusteellisuudella tarkoitetaan sitä, että dokumentin sisältö kuvataan mahdollisimman tarkasti erilaisilla asiasanoilla. Tarkkuudella taas tarkoitetaan sitä, miten tarkkoja ja kapeita käsitteitä asiasanoiksi valitaan. Tätä kautta asiasanojen valinnassa voidaan karkeasti erotella kaksi erilaista tapaa: syvällinen ja pinnallinen asiasanoitus. Näistä ensin mainittu (engl. *deep indexing*) tarkoittaa sitä, että sekä perusteellisuus ja tarkkuus pyritään maksimoimaan. Tällöin aineistot kuvaillaan mahdollisimman laajasti mahdollisimman tarkoin termein, jolloin hakutulokset ovat tarkkoja, mutta ongelmaksi voi muodostua asiasanojen valinta. Pinnallinen asiasanoitus (engl. *shallow indexing*) taas tarkoittaa sitä, että aineistoja kuvaillaan muutamalla yleisellä termillä, jolloin hakutulokset voivat heikentyä, mutta itse asiasanoitusprosessi saadaan kevyemmäksi. [SM86]

Valittiinpa asiasanoitusmenetelmäksi syvällinen tai pinnallinen tapa, on tärkeää, että valittua tapaa noudatetaan yhdenmukaisesti aineistojen ja asiasanoitusten laatijien välillä. Toisaalta jos asiasanoituksen perustana käytetään jotain ennaltasovittua sanastoa, voidaan hyödyntää siitä löytyviä käsitesuhteita, jolloin asiasanojen tarkkuuden merkitys pienenee. Semanttisessa webissä voidaan hyödyntää formaalisti määriteltyjä ontologioita sanastoina. Tällöin asiasanoiksi kannattaneekin valita mahdollisimman tarkkoja käsitteitä, koska haku voidaan kohdistaa myös yläluokan käsitteisiin. Esimerkiksi jos aineistosta haetaan hakusanalla *koira*, ja tiettyyn dokumenttiin on liitetty asiasana *bokseri*, palautetaan tämä dokumentti, koska tiedämme sanaston perusteella bokserin olevan koiran alakäsite. Toisaalta taas, jos dokumentti olisi asiasanoitettu laajemmalla *koira*-käsittellä, ei *bokseri*-haku enää löytäisi kyseistä dokumenttia. Ongelma tosin ratkeaisi tekemällä hakusanan laajennus yläkäsitteeseen, mutta kuten Bilotti [Bil04] toteaa, ei hakusanojen laajentaminen välttämättä paranna hakutuloksia ja on ongelmallista.

3.5 Asiasanoitus semanttisessa webissä

Yksinkertaisin tapa käyttää asiasanoitusta on liittää dokumentin yhteyteen pelkkä merkkijonolista käytetyistä asiasanoista. Tällainen asiasanoitus on

helppo tehdä ja ihmiset ymmärtävät sitä helposti. Ongelmana on kuitenkin se, että koneet eivät osaa tulkita näitä asiasanoja ja niiden merkitystä. Jotta koneet ymmärtäisivät asiasanoituksia, ja jotta semanttisen webin visio voisi toteutua, tulee asiasanat sitoa koneiden ymmärtämiin ontologioihin. Ontologiasidonnaisuus on tärkeää semanttisen yhteensopivuuden kannalta [HVK+05]. Ontologiasidonnaisesta asiasanoituksesta käytetään tässä raportissa termiä *semanttinen asiasanoitus*.

Ontologioihin pohjautuva semanttinen annotointi voidaan jakaa vapaaseen ja skeemapohjaiseen annotointiin [SDWW01]. Vapaassa annotoinnissa käytettävät ontologiat ovat kuvauksia jonkin tietyn aihealueen sanastosta tai ne voivat olla yleissanastoja. Tässä sanan *vapaus* käyttö tulee siitä, ettei käytettäviä asiasanoja rajoiteta mitenkään ehkä lukumäärää lukuunottamatta. Skeemapohjaisessa annotoinnissa käytetään jotain ontologiaa – *annotaatio-ontologiaa*, joka kertoo, mitä annotoitavasta dokumentista halutaan kuvailla. Termit semanttinen asiasanoitus ja annotointi ovat läheisiä ja voivat mennä sekaisin. Tässä työssä semanttisella asiasanoituksella tarkoitetaan vapaata, johonkin yleiseen sanastoon perustuvaa semanttista annotointia.

Annotaatio-ontologian käytössä on muun muassa se etu, että sen avulla voidaan soveltaa helpommin näkymäpohjaista hakua [HSV04], koska dokumentista on tällöin kuvailtu samat, annotaatio-ontologiassa määritellyt asiat. Toisaalta, jos dokumenttien sisällön aihepiiri ei ole rajattu, voi annotaatio-ontologian käyttäminen olla keinotekoista, koska dokumenteissa ei ole tunnistettavissa yhteisiä piirteitä. Skeemapohjaista annotointia kannattaa kuitenkin harkita vapaan asiasanoituksen tilalle, jos sovelluksen aineisto on jonkin verran homogeenistä ja siitä on tunnistettavissa yhteisiä piirteitä.

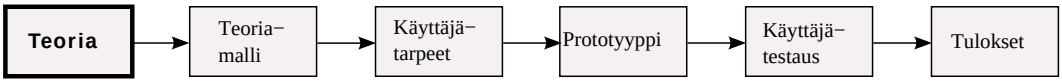
Semanttisessa asiasanoituksessa voidaan dokumentteihin käsitteiden lisäksi liittää *ilmentymiä*, joilla tarkoitetaan käsitteistä luotuja yksilöitä. Esimerkiksi *Santahamina* voisi olla käsitteen *varuskunnat* ilmentymä ja *Garfield* käsitteen *sarjakuvahahmot* ilmentymä.

Semanttisen asiasanoituksen selkeä etu on se, että asiasana on sidottu johonkin ontologiseen formaalisti määriteltyyn käsitteeseen. Esimerkiksi käytettäessä asiasanaa *johtaminen* ei perinteisessä asiasanoituksessa voida päätellä (paitsi ehkä muista asiasanoista ja asiasanoitettavan dokumentin sisällöstä) mistä johtamisesta on kyse: sähkön johtaminen, joukkojen johtaminen, kilpailussa johtaminen ja niin edelleen. Kun asiasanoituksen pohjana käytetään ontologisia käsitteitä, voidaan asiasana sitoa tarkasti tiettyyn johtamiseen, ja edellä kuvattu ongelma poistuu. Tällainen tarkkuus asiasanoituksessa voidaan nähdä kuitenkin myös huonona puolena, jos asiasanoittaja ei halua tai osaa käyttää asiasanan eksplisiittistä merkitystä. Esimerkiksi asiasana *lapset* voi tarkoittaa

esimerkiksi lapsia ikäryhmänä, lapsia sosioekonomisessa roolissa tai lapsia vanhempiensa jälkeläisinä. Usein voi olla, että asiasanoittaja haluaisi käyttää asiasanaa sellaisenaan sitomatta sitä tarkasti tiettyyn ontologiseen käsitteeseen.

Luku 4

Ongelmanratkaisu tapauksiin perustuvan päättelyn avulla



Edellisessä kappaleessa kysymys-vastauspalveluita tarkasteltiin asiasanoittamisen näkökulmasta; miten kysymys-vastausaineistoa kannattaa asiasanoittaa, jotta haut aineistoon olisivat mahdollisimman tehokkaita. Tässä kappaleessa pureudutaan problematiikkaan, joka toistuu erilaisissa kysymys-vastauspalveluissa: miten voidaan hyödyntää koneellista päättelyä ja olemassa olevaa kysymys-vastausarkistoa siten, että kysymykseen vastaamista ei tarvitsisi tehdä täysin manuaalisesti.

4.1 Taustaa ongelmanratkaisujärjestelmistä

Ongelmanratkaisulla tarkoitetaan menetelmiä, joilla koneellista päättelyä ja järjestelmän käyttämää aineistoa hyväksikäyttäen ratkaistaan käyttäjän ongelma. Erilaisia ongelmanratkaisujärjestelmiä on monenlaisia, mutta niillä kaikilla on sama perusperiaate: käyttäjällä on jokin kysymys, ongelma, johon hän etsii vastausta. Yksinkertainen tapa rakentaa ongelmanratkaisujärjestelmä kysymys-vastausaineistosta on syöttää se relaatio- tai oliotietokantaan, ja tarjota käyttäjälle yksinkertainen tekstihaku aineistoon. Joissain tapauksissa tällainen ratkaisutapa voikin olla täysin riittävä, eikä hienostuneempia menetelmiä tarvita. Jos ongelma-alue on laaja, tai jos aineistoa on runsaasti, on

syytä kuitenkin soveltaa jotain systemaattista menetelmää ongelmien ratkaisemiseen.

4.1.1 Tiedonhakujärjestelmät

Yksinkertaisissa tietokantasovelluksissa menetelmät tekstin hakuun ovat kovin rajoittuneita. Esimerkiksi synonyymejä ei tunnisteta, ja yleensä pieni kirjoitusvirhe hakumerkkijonossa tarkoittaa, ettei hakutuloksia löydetä. Ihmisen voi olla myös hankala kuvata ongelmiaan, ja usein eri ihmiset käyttävät samasta käsitteestä eri termiä. Esimerkiksi Furnas ym. [FLGD87] toteavat, että yleisessä tapaukssa ihmiset valitsevat samasta käsitteestä saman termin alle 20% todennäköisyydellä.

Yksinkertainen tietokantahaku on oikeastaan datan eikä tiedon hakemista. Jotta voidaan puhua tiedonhakujärjestelmästä (engl. *Information Retrieval System*), pitää järjestelmän jollain tapaa "tulkita" aineistoa [BYRN99]. Tiedonhakujärjestelmät voidaankin nähdä eräänlaisina hienostuneempina versioina yksinkertaisista tietokantasovelluksista. Watson [Wat98] luonnehtii tiedonhakujärjestelmiä siten, että käyttäjä usein syöttää kysymyksen luonnollisella kielellä, haun taustalla on jokin synonyymisanasto, ja usein hakuja tehdään valtavaan tekstimäärään.

4.1.2 Sääntöihin perustuvat asiantuntijajärjestelmät

Seuraavaksi askeleeksi tiedonhakujärjestelmistä voidaan nähdä sääntöihin perustuvaa päättelyä (engl. *Rule-based Reasoning*) käyttävät järjestelmät. Niissä ongelma-alueesta luodaan sääntöjä, joihin ongelmanratkaisu perustuu. Esimerkiksi *jos asiakas on alle 60-vuotias, anna alennus* ja *alennuksen suuruus on 60 euroa*. Tällöin ongelmanratkaisujärjestelmä koostuu suuresta joukosta tällaisia sääntöjä, joiden perusteella ratkaisu ongelmaan päätellään [Wat98].

Sääntöihin perustuvissa asiantuntijajärjestelmissä on se haittapuoli, että ne vaativat yleisesti ottaen sen, että jokin taho ylläpitää sääntöjä, joihin järjestelmä perustuu. Usein voi olla niin, että tämä taho on muutenkin työllistetty, jolloin sääntöjen ylläpito jää taka-alalle. Lisäksi, jos ongelma-alue ei ole tarkkaan rajattu, voi sääntöjen keksiminen olla vaikeaa, eikä riittävän kattavaa säännöstöä ole mahdollista laatia. Näissä tapauksissa sääntöihin perustuvia asiantuntijajärjestelmiä ei kannatta käyttää. Toisaalta jos ongelma-alue on luonteeltaan muuttumaton ja uusia sääntöjä ilmaantuu harvoin tai ei ollenkaan, voi sääntöihin perustuva järjestelmä olla täysin riittävä.

4.2 Tapauksiin perustuva päättely

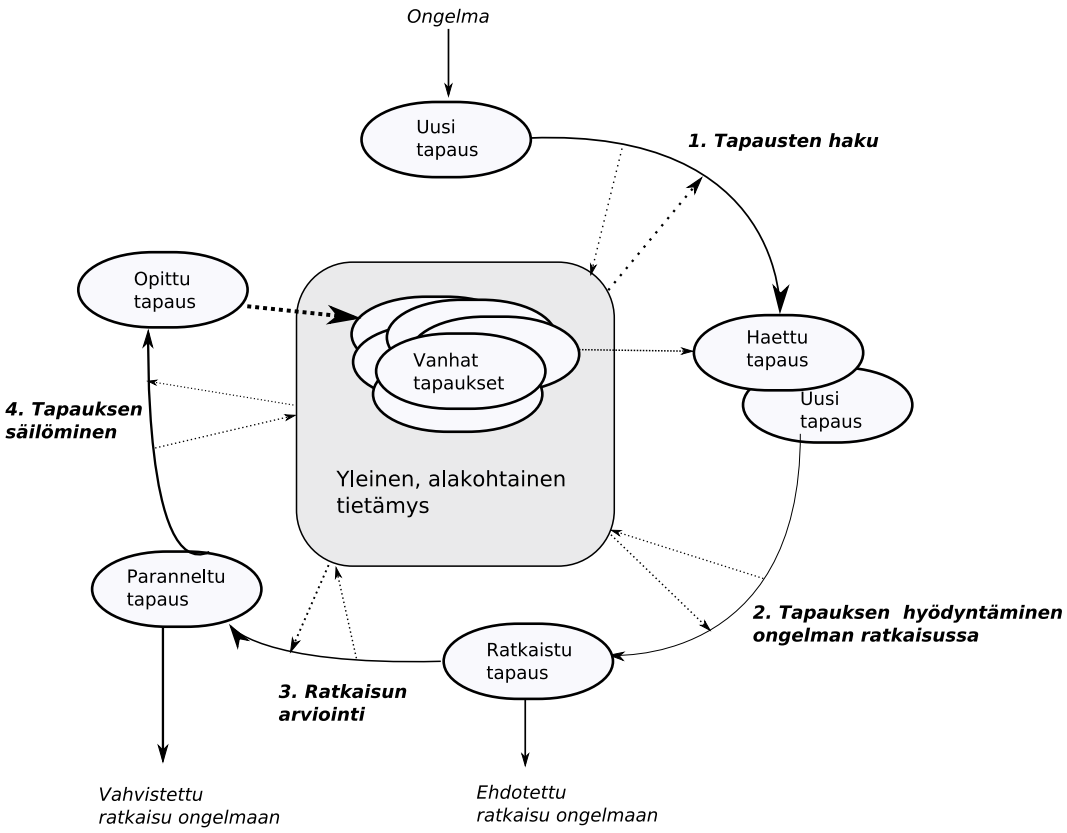
Tapauksiin perustuva päättely (engl. *Case-based Reasoning*) on ongelmanratkaisumenetelmä, joka perustuu ihmisen luontaiseen tapaan ratkaista ongelmia. Siinä ratkaisuja uusiin ongelmiin etsitään vanhojen, aiemmin ratkaistujen ongelmien perusteella. Ongelmasta ja sen ratkaisusta tehtyä kuvausta kutsutaan *tapaukseksi*. Tietovarastoa, johon tapaukset on talletettu, kutsutaan *tapauskannaksi*.

Ihmisten pulmat tapaavat toistua, kuten esimerkiksi UKK-listauksien suosiosta voidaan päätellä. Ilmiö voidaan nähdä myös muualla kuin ihmisten pulmissa. Esimerkiksi eräässä teknisessä neuvontapistepalvelussa 60 prosenttia ongelmista aiheutui samankaltaisesta laiteviasta [Sim92]. Tapauksiin perustuva päättely on siis ongelmanratkaisumenetelmä, joka hyödyntää ihmisten pulmien ja erilaisten ilmiöiden samankaltaisuutta ja toistuvuutta. Selkeyden vuoksi tässä raportissa joissain kohdissa tapauksiin perustuvasta päättelystä käytetään englanninkielistä lyhennettä *CBR*.

Karkeasti jaoteltuna tapauksiin perustuva päättely koostuu neljästä askeleesta [Wat98], joskin menetelmä voidaan jakaa myös viiteen askeleeseen (esimerkiksi [CRCC96, FS03]), jolloin tapausten esittäminen on menetelmän ensimmäinen askel. Nämä askeleet muodostava niin sanotun *CBR-syklin*:

1. Tapausten esittäminen
2. Ongelmaa muistuttavien tapausten haku
3. Haun perustella löydettyjen tapausten hyödyntäminen ongelman ratkaisussa
4. Ratkaisun arviointi tarvittaessa
5. Ratkaisun säilöminen uutena tapauksena

Kuvassa 4.1 on esitetty tapauksiin perustuvan päättelyn vaiheet. Huomattavaa on se, että olemassa olevia tapauksia voidaan käyttää myös muissa kuin ensimmäisessä hakuaskeleessa. Olemassa olevat tapaukset on sijoitettu yleisen tietämyksen (engl. *general knowledge*) sisään. Päättely voi siis varsinaisten säilöttyjen tapausten lisäksi perustua myös yleiseen alakohtaiseen tietämykseen. Esimerkiksi lääketieteellisessä sovelluksessa, jossa tapauksen muodostaa kunkin potilaan sairaskertomus, yleiseen tietämykseen voi sisältyä esimerkiksi tietämys anatomian säännöistä.



Kuva 4.1: Tapauksiin perustuvan päättelyn neljä askelta [AP94]

Kuvassa ei ole havainnollistettu syklin ulkopuolella olevaa ihmistä, jolle ehdotetaan ratkaisua ongelmaan, ja joka vahvistaa ratkaisun.

4.2.1 Tapauksiin perustuvat päättelytavat

Tapauksiin perustuva päättely ei ole mikään yksittäinen tekniikka, vaan pikemminkin lähestymistapa ongelmien ratkaisuun, ja kunkin askelen toteutustapa on sovelluskohtainen. Hyvin usein ihmisellä on olennainen osa tapauksiin perustuvan päättelyn askelissa, ja vain harvoin kaikki vaiheet tehdään koneellisesti. Tapauksiin perustuvaa päättelyä voidaan pitää oikeastaan useamman erilaisen lähestymistavan kokoavana käsitteenä. Esimerkiksi Aamodt ja Plaza [AP94] jakavat sen seuraaviin alamenetelmiin:

Malleihin perustuva päättely (engl. *Exemplar-based reasoning*). Tässä

menetelmässä tarkoituksena on löytää pelkästään luokka uudelle tapaukselle. Esimerkiksi uudelle elokuvalle annetaan ikäluokitus tapauskannassa olevien elokuvien attribuuttien sekä niille annettujen luokitusten perusteella.

Yksilöihin perustuva päättely (engl. *Instance-based reasoning*) [AKA91]. Menetelmä on malleihin perustuvan päättelyn erikoistapaus, jossa luokitteluarvioita tehdään pelkästään tiettyjen yksilöiden, instanssien, perusteella ilman yleistä tietämystä. Tälle tekoälysovelluksissa usein käytetylle menetelmälle ominaista on, että CBR-sykli on täysin automaattinen. Esimerkiksi edellisen kohdan elokuvan ikäluokitustehtävässä luokitus annettaisiin pelkästään muiden elokuvien pohjalta tukeutumatta yleiseen tietämykseen, kuten lainsäädäntöön ja ilman, että elokuvatarkastaja vahvistaa päätellyn ikäluokituksen.

Muistiin perustuva päättely (engl. *Memory-based reasoning*). Tässä menetelmässä tapauskanta nähdään suurena muistina. Painotus on tapauskannan järjestelyssä ja tehokkaassa käsittelyssä sekä siinä, miten tapauskantaan päästään käsiksi. Esimerkkinä voisi olla jokin jättiläismäinen potilastietokanta ja sen perusteella tehtävät päättelyt uudelle potilaalle.

Tapauksiin perustuva päättely (engl. *Case-based reasoning*). Menetelmällä on hieman harhaanjohtavasti sama nimi kuin tapauksiin perustuvalla päättelyllä yleisessä merkityksessä. Tässä alamenetelmässä painotetaan kuitenkin yleisten tietämyksen käyttöä sekä sitä, että tapaukset ovat sisäiseltä rakenteeltaan joissain määrin monimutkaisia, ja että järjestelmä pystyy osallistumaan löydettyjen tapausten hyödyntämiseen. Esimerkkinä voisi olla järjestelmä, joka ehdottaa uusia ruokareseptejä olemassa olevien reseptien perusteella. Yleisenä tietämyksenä tällöin on yleiset ruuanlaiton lainalaisuudet erilaisten ainesten ja makujen yhdistämisestä ja ruoka-aineiden ravintoarvoista.

Yhdenmukaisuuteen perustuva päättely (engl. *Analogy-based reasoning*). Tämä menetelmä on hyvin lähellä tyypillistä CBR-lähestymistapaa, joskin erona on se, että toisin kuin CBR:ssä yleensä, menetelmässä käytetään useita ongelma-alueita tavoitteena löytää yhdenmukaisuuksia. Esimerkiksi terveydenhuollon alalla voitaisiin käyttää biologisia tapauksia, ja edellisen kohdan esimerkissä reseptiehdotuksia sovellettaisiin vaikkapa eläinten ruokintaan.

Etenkin jos yksittäiset tapaukset ovat monimutkaisia, olennaista on myös se, miten tapaukset esitetään. Finnie [FS03] kritisoikin tapauksiin perustuvan

päättelyn eri teorioita siitä, että usein tapausten esittäminen unohdetaan. Seuraavassa tapauksiin perustuvan päättelyn vaiheisiin ja tapausten esittämiseen pureudutaan hieman tarkemmin.

4.2.2 Tapauksiin perustuvan päättelyn askelet

Tapausten esittäminen

Tapaukset ovat nimensä mukaisesti tapauksiin perustuvan päättelyn ydin. Siksi tapauskannan käsittely pitää olla tehokasta ja nopeaa. Aamodtin ja Plazan [AP94] mukaan tässä kohdassa pitää päättää seuraavat kohdat:

- Mitä ylipäänsä tapauksesta halutaan kuvata?
- Minkälainen rakenne on sopiva kuvauksen tekemiseen?
- Miten tapausmuisti organisoidaan ja indeksoidaan tehokasta hakua varten?
- Miten tapaukset integroidaan yleiseen tietämykseen?

Semanttisessa webissä ontologiat tarjoavat hyvän lähtökohdan näihin kohtiin. Niillä voidaan luoda malli tapauksesta, ja jos yleinen tietämys on myös annotoitu ontologiapohjaisesti, on tapausten integrointi muihin lähteisiin helppoa.

Chen ja Wu [CW03] ovat kehittäneet CaseML:n, joka on RDF-pohjainen kieli tapausten esittämiseen semanttisessa webissä. Kieltä voidaan käyttää yleisenä kehyksenä tapausten kuvaamiseen, ja se sallii tapausten integroinnin aihepiirikohtaisiin ontologiapohjaisiin sanastoihin.

Tapausten haku

Aamodtin ja Plazan [AP94] mukaan yksinkertaistettuna tapausten haku alkaa ongelman kuvauksella ja päättyy siihen, kun parhaiten ongelmaa vastaava tapaus on löydetty. Falkman [Fal03] taas jakaa ensimmäisen askelen erikseen hakuun ja hakutulosten järjestämiseen. Falkmanin jaossa otetaan enemmän huomioon ihmisen osallistuminen sykliin hakutulosten vahvistajana, kun taas Aamodtin ja Plazan jaossa hakutulosten järjestäminen ja luokittelu on järjestelmän tehtävä, ja tuloksena on vain yksi, paras tapaus.

Tarkemmin tarksteltuna tapausten haku voidaan jakaa seuraaviin alakohtiin [AP94]:

- Uuden tapauksen muodostaminen ja sen piirteiden tunnistaminen ongelmakuvauksen perusteella. Jos ongelman kuvaus on puutteellinen, voidaan käyttäjää pyytää täydentämään puuttuvia tapauksen attribuutteja.
- Uutta tapausta vastaavien tapausten etsiminen tapauskannasta. Hakutavat ovat hyvin sovelluskohtaisia.
- Parhaan tapauksen valinta edellisessä kohdassa löydetyistä samankaltaisista tapauksista. Usein tässä vaiheessa löydetyt tapaukset järjestetään jonkin kriteerin perusteella.

Näissä askelissa haku on oikeastaan jaettu kahteen askeleeseen: ensin haetaan väljillä hakukriteereillä joukko samankaltaisia tapauksia ja vasta tämän jälkeen tästä joukosta valitaan paras tai parhaat tapaukset.

Löydettyjen tapausten hyödyntäminen

Yksinkertaisimmillaan löydetyn tapauksen hyödyntäminen voi olla sitä, että vanha tapaus kopioidaan uudeksi tapaukseksi. Esimerkiksi uudelle elokuvalle ikäluokitus kopioidaan suoraan elokuvaa eniten vastaavasta elokuvasta. Useammin kuitenkin vanhaa tapausta ei kopioida sellaisenaan, vaan siitä *mukautetaan* uusi tapaus. On hyvin sovelluskohtaista, miten vanhan tapauksen mukauttaminen tapahtuu. Varsinaisen muuntamisen eli vanhan tapauksen attribuuttien käyttämisen lisäksi voidaan myös hyödyntää menetelmää, jolla vanha tapaus luotiin. Menetelmän hyödyntäminen edellyttää, että tapaukseen on säilötty metatietoa siitä, miten tapaus on luotu.

Ongelmanratkaisujärjestelmissä on tärkeää, että käyttäjälle näytetään järkeily järjestelmän tekemän päättelyn takana. Erityisen tärkeää tämä on silloin, kun järkeily on tehty monimutkaisen päättelyn perusteella, eikä esimerkiksi perustuen yksinkertaiseen luokkien perintäsuhteeseen [MPS98]. Järkeilyn perustelu on myös tärkeää silloin, kun ongelmaan ehdotettu ratkaisu on kriittinen esimerkiksi ihmishenkien kannalta, kuten lääketieteessä. Tapauksiin perustuvassa päättelyssä tämä pitää ottaa huomioon siten, että jos järjestelmä tekee tapausten hyödyntämisaskeleessa monimutkaista mukauttamista tapaukselle, pitää mukautus perustella.

Ongelmanratkaisujärjestelmissä on myös tärkeää, että ratkaisuissa ei esitetä liikaa tietoa, vaan siitä karsitaan käyttäjälle irrelevantti osuus pois [MPS98]. Tässä on huomioitavaa se, että tiedon relevanttius on aina kontekstiriippuvaista, eli jossain yhteydessä turha tieto voi olla olennaista tietoa toisaalla.

Ratkaisun arviointi

Edellisen askeleen lopputulemana CBR-järjestelmä antaa ratkaisuehdotuksen ongelmaan. Jos CBR-sykli on täysin automaattinen, ratkaisun arviointi on jo tehty edellisessä askeleessa. Useimmissa sovelluksissa järjestelmän ehdottama ratkaisu täytyy kuitenkin jotenkin evaluoida. Tämä voi tarkoittaa esimerkiksi ratkaisun kokeilua oikeassa elämässä. Reseptejä ehdottavassa CBR-järjestelmässä tämä voisi tarkoittaa sitä, että järjestelmän ehdottamalla reseptillä tehtyä ruokaa maistetaan. Sovelluksista riippuen ratkaisun arviointi voi viedä kauankin, jolloin ratkaisuehdotus voidaan jo säilöä tapauskantaan ratkaistuna tapauksena, mutta silloin tapaukseen pitää lisätä jokin merkki, ettei sitä vielä ole arvioitu.

Arviointiaskeleessa myös korjataan järjestelmän ehdottamaa ratkaisua. Tässä voidaan kertoa järjestelmälle, mikä ratkaisussa oli virheellistä ja "palauttaa" se edelliseen askeleeseen, tai sitten ratkaisun korjaaminen tehdään manuaalisesti.

Ratkaisun säilöminen

Tapauksiin perustuva päättely perustuu oikeastaan viimeiseen askeleeseen, eli siihen, että ratkaisu ongelmaan säilötään tapauskantaan uutena tapauksena. Yksinkertaisissa sovelluksissa tämä askel voi olla triviaali, mutta etenkin monimutkaisissa sovelluksissa pitää miettiä muun muassa sitä, mitä ratkaisumenetelmästä säilötään uuden tapauksen yhteyteen, ja miten uusi tapaus integroidaan tapauskantaan ja yleiseen tietämykseen.

4.2.3 Katsaus CBR-järjestelmiin

Tapauksiin perustuva päättely tuntuu hyvin intuitiiviselta lähestymistavalta ongelmanratkaisuun, ja menetelmä istuu hyvin kysymys-vastaustyyppisiin järjestelmiin; onhan kysymys helposti nähtävissä ongelman kuvaukseksi ja vastaus ratkaisuksi siihen. Tällöin kysymys-vastauspari muodostaa CBR:n *tapauksen*. Myös Goker ja Roth-Berghofer [GRB99] toteavat, että CBR:ää voidaan soveltaa kysymyksiin ja vastauksiin perustuvissa tukikeskuspalveluissa.

Käyttämällä CBR:ää organisaatio vahvistaa yleistä tietämystään ja tehostaa toimintaansa vähentämällä kysymyksiin vastaamiseen kuluvaa aikaa. Kai ym. [CRCC96] näyttää, että tukihenkilöt, jotka ratkaisevat ongelman CBR-järjestelmän avulla, muistavat ratkaisun pitempään. Tämä johtuu siitä, että heille tulee vaikutelma, että he olisivat ratkaisseet ongelman itse, vaikka itse asiassa ratkaisu olisi haettu ja mahdollisesti mukautettu tapauskannasta.

Tapauksiin perustuva päättely on perinteisesti ollut vahvasti esillä lääketieteellisissä sovelluksissa [Fal03, SG01, Bic04]. Tämä johtunee ainakin osin siitä, että CBR tukee hyvin sitä tapaa, jolla lääkärit toimivat; hoitopäätökset tehdään usein edellisten hoitopäätösten tuloksiin ja lääkärin hoitokokemuksiin nojautuen. Usein myös tapausten esittäminen on helppoa, esimerkiksi potilaskuvauksista saadaan helposti skalaarisia arvoja, kuten verenpaine, kolesteroliarvot ja niin edelleen.

Bichindaritz [Bic04] esittelee Mémoiren, joka on semanttisen webin tekniikoihin perustuva kehys tapauskantojen jakamiseen ja tapauksiin perustuvan päättelyn käyttämiseen lääketieteessä ja biologiassa. Artikkelissa todetaan, että semanttisen webin edut näkyvät siinä, että eri alojen sovellukset voidaan liittää vaivatta yhteen yhteisten ontologioiden avulla. Bichindaritz arvioi semanttisen webin tekniikoiden sopivan hyvin yhteen tapauksiin perustuvan päättelyn kanssa.

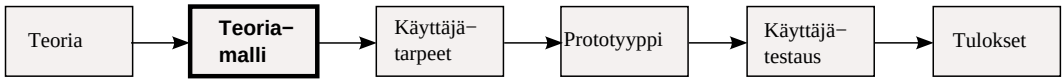
4.3 Ongelmanratkaisumenetelmän valinta

Aina ongelmanratkaisujärjestelmiä ei voida lokeroida suoraan tiettyyn tyyppiin. Esimerkiksi Golding ja Rosenbloom [GR96] ovat tutkineet, miten sääntöihin ja tapauksiin perustuvaa päättelyä voisi yhdistää. Menetelmässä ongelmien ratkaisu tehdään ensisijaisesti asiantuntijoiden laatimien sääntöjen mukaan, ja tapausten roolina on täydentää ongelma-alueita, johon säännöt eivät ulotu, sekä vahvistaa tai heikentää säännöistä saatuja ratkaisuja. Tämä soveltuukin hyvin ongelma-alueisiin, jotka ovat suhteellisen hyvin tunnettuja, mutta kuitenkin sisältävät jonkin verran tuntematonta tietoa.

Oleellista ongelmanratkaisumenetelmää valittaessa on muistaa, että ei välttämättä kannata lukkiutua tiettyyn menetelmään, vaan valita parhaita puolia eri ratkaisutavoista. Tärkeää on myös muistaa, ettei ongelmanratkaisu yleensä lopu käyttäjän ensimmäisen haun tuloksiin, vaan hakutulosten perusteella käyttäjälle pitää antaa mahdollisuus rajata tai laajentaa hakuaan. Haku pitääkin nähdä prosessina, eikä yksittäisenä toimenpiteenä.

Luku 5

Malli ontologiapohjaisesta kysymys-vastauspalvelusta



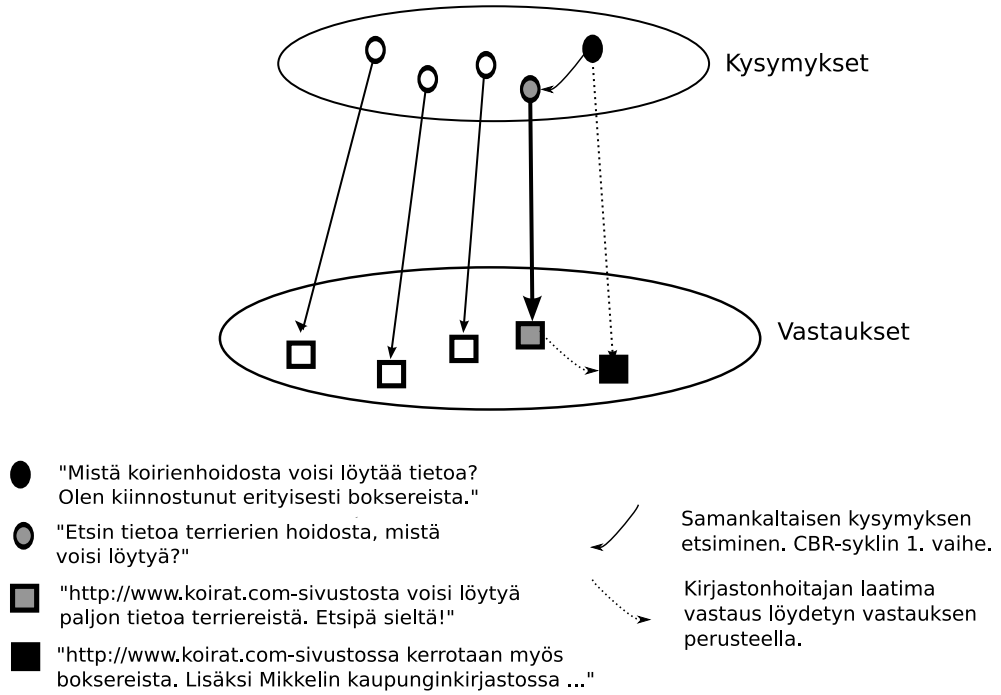
Tässä kappaleessa kootaan yhteen edellisissä kappaleissa esiteltyä teoriaa ja luodaan teoreettinen synteesi, *malli*, siitä, miten semanttisen webin tekniikoita ja tapauksiin perustuvaa päättelyä voi käyttää kysymys-vastauspalveluissa. Mallissa yhdistetään semanttinen asiasanoitus tapauksiin perustuvan päättelyn askeliin, ja se soveltuu niin uusien kysymys-vastauspalveluiden luontiin kuin vanhojen palveluiden muuntamiseen ontologiseen muotoon. Malli keskittyy kysymyksiin vastaajan rooliin palveluissa.

Ensin luodaan yleiskuva CBR:ää hyödyntävästä kysymys-vastauspalvelusta, minkä jälkeen esitellään vaatimukset ja tarvittavat valmistelut mallin luomista varten. Lopuksi käydään läpi, miten malli osallistuu CBR-syklin askeleisiin.

5.1 Yleiskuva CBR:ää hyödyntävästä kysymys-vastauspalvelusta

Kuvassa 5.1 on esimerkki siitä, miten tapauksiin perustuvaa päättelyä voitaisiin käyttää kysymys-vastauspalvelussa. Esimerkiksi on otettu Kysy kirjastonhoitajalta -tyyppinen aineisto. Kuvassa nähdään, miten käyttäjä on esittänyt

kysymyksen liittyen boksereiden hoitoon. CBR:n ensimmäistä askelta (Ongelmaa muistuttavien tapausten haku) käyttäen löydetään toinen koirienhoitoon, mutta hieman eri rotuun liittyvä kysymys. Kirjastonhoitaja käyttää tämän kysymyksen vastausta pohjana laatiessaan vastauksen käyttäjän kysymykseen (toinen ja kolmas askel). Lopuksi uusi, muokattu vastaus tallennetaan vastaustietokantaan, josta se on käytettävissä tulevaisuutta varten (neljäs askel).



Kuva 5.1: Esimerkki CBR:n soveltamisesta kysymys-vastauspalvelussa

Kuten aiemmin esitetty, tapauksiin perustuva päättely ei ole mikään yksittäinen tekniikka vaan voidaan jakaa useampaan alamenetelmään. Mutta mitä näistä alamenetelmästä kysymys-vastauspalveluissa voisi soveltaa? Alamenetelmistä tapauksiin perustuva päättely (katso 4.2.1), ei sellaisenaan sovi, koska kysymys-vastaustyyppisessä aineistossa tapausten kuvaukset harvemmin ovat monimutkaisia (kysymysteksti ja ehkä vähän metatietoa), ja kun käytetään luonnollista kieltä, ei vastausten muodostaminen automaattisesti kuulosta vaatimansa vaivan arvoiselta.

Kysymys-vastauspalveluissa olennaisinta lienee se, miten kaksi kysymystä tunnistetaan samankaltaisiksi ja toisaalta myös, miten löydetyt samankaltaiset kysymykset järjestetään relevanssin mukaan. Problematiikka painottuu siis CBR-syklin ensimmäiselle askeleelle, ja siinä on sekä muistiin perustuvan, yksilöihin

perustuvan että tapauksesta riippuen myös yhdenmukaisuuteen perustuvan päättelyn tunnusmerkkejä.

5.2 Mallin vaatimukset ja valmistelu

Malli on riippuvainen kahdesta ulkoisesta komponentista:

Sanasto-ontologia Asiasanoituksen pohjana käytetään laajaa yleistä sanasto-ontologiaa. Käytettävä ontologia voi kuvata myös jonkin tietyn aihealan tarkemman sanaston. Myös useampia ontologioita voidaan käyttää.

Semanttinen tiedoneristäjä (STE) Semanttinen tiedoneristäjä etsii luonnollisen kielen tekstistä ontologisia käsitteitä edellä kuvatusta sanasto-ontologiasta tai -ontologioista.

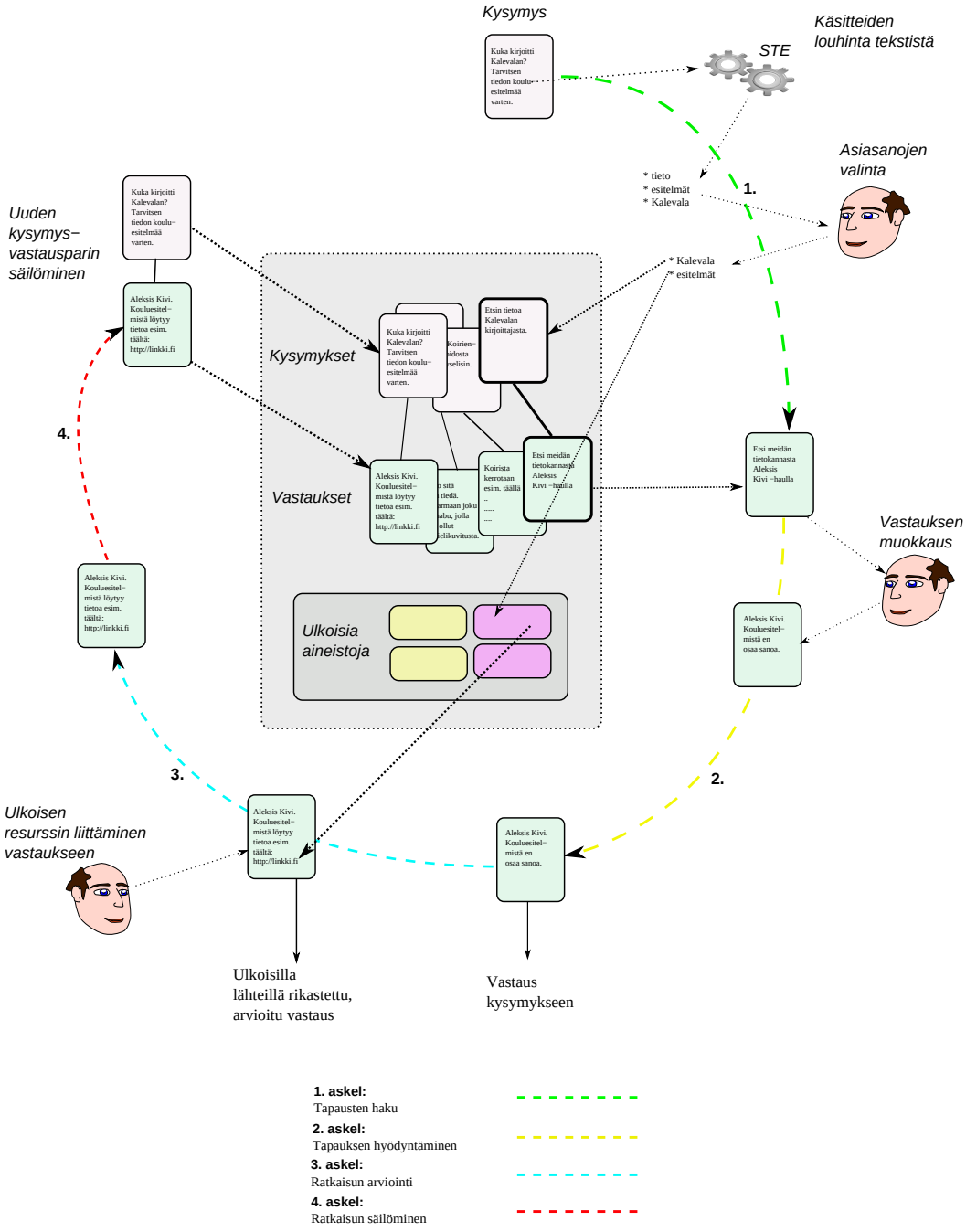
Mallin käyttämät ulkoiset aineistot muunnetaan ontologiamuotoon kappaleessa 2.2.3 esiteltyjen periaatteiden mukaan. Ulkoisten aineistojen asiasanat ja annotaatiot liitetään sanasto-ontologian käsitteisiin mahdollisuuksien mukaan.

5.3 CBR-syklin askeleet

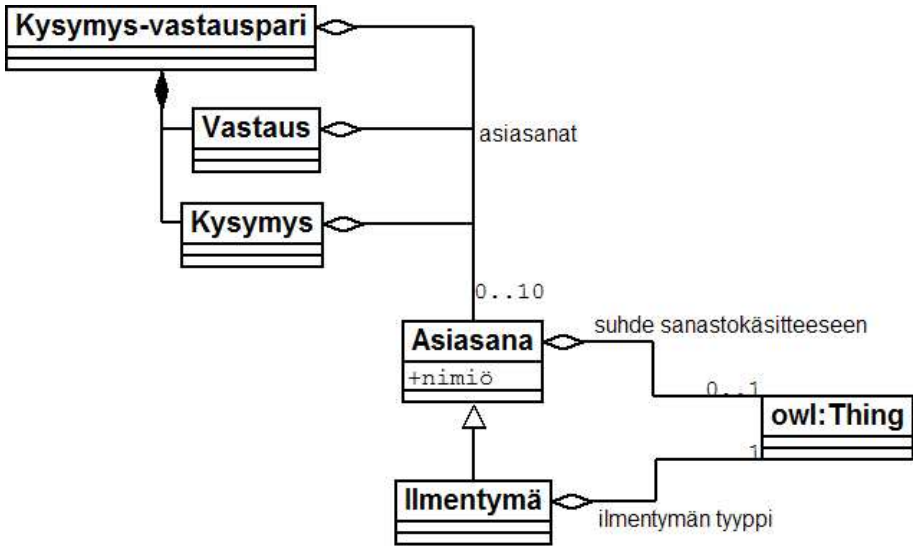
Tapauksiin perustuvaa päättelyä käsittelevässä kirjallisuudessa painotetaan usein, miten ihminen usein osallistuu päättelyketjun askeliin, ja miten harvemmin koko CBR-sykli tapahtuu koneellisesti. Kuitenkin erilaisista CBR-sykliä esittävistä mallikaavioista puuttuu ihmisen osuuden mallinnus. Kuvassa 5.2 on esitetty malli, jossa semanttinen asiasanoitus on nivottu yhteen tapauksiin perustuvan päättelyn askelten kanssa kysymys-vastauspalvelussa, ja siinä on myös havainnollistettu, miten tässä mallissa ihminen osallistuu tapauksiin perustuvan päättelyn askeliin.

5.3.1 Tapausten esittäminen

Tapausten esittämistä varten laaditaan kysymys-vastausontologia, jota on havainnollistettu luokkakaaviona kuvassa 5.3. Ontologiassa luokka *Kysymys-vastauspari* pitää sisällään nimensä mukaisesti *Kysymyksen* ja *Vastauksen*.



Kuva 5.2: Semanttinen asiasanoitus yhdistettynä tapauksiin perustuvan päättelyn askeliin



Kuva 5.3: Luokkakaavio tapausten esittämiseen käytettävästä ontologiasta

Kaaviossa *Asiasanoja* ja *Ilmentymiä* voi olla sekä kysymyksellä, vastauksella että myös kysymys-vastausparilla. Tarkoituksena on, että uutta aineistoa luotaessa asiasanat ja ilmentymät liitetään erikseen sekä kysymykseen että vastaukseen. Näin samankaltaisia kysymyksiä etsittäessä haku kohdistetaan kysymykseen ja sen asiasanoihin. Se, että myös kysymys-vastausparilla voi olla asiasanoja ja ilmentymiä, on tehty vanhoja aineistoja varten. Niissä asiasanat on yleensä liitetty koko kysymys-vastauspariin. Kaaviossa on myös esitetty, että asiasanoja olisi maksimissaan kymmenen.

Kullakin asiasanalla on suhde yhteen tai useampaan ontologiseen käsitteeseen. Kaaviossa tätä on kuvattu siten, että luokalla *Asiasana* on suhde *owl:Thing*:in, joka on kaikkien ontologisten käsitteiden yläluokka. Myös luokalla *Ilmentymä* on suhde *owl:Thing*:in. Tällä suhteella ilmaistaan minkä luokan ilmentymästä on kyse. Asiasanalla ei välttämättä ole suhdetta ontologiseen käsitteeseen, mutta ilmentymällä on aina oltava tyyppi. Kaaviossa *Ilmentymä* on *Asiasanan* alaluokka, koska tavallaan ilmentymät ovat asiasanojen erikoistapauksia.

Tarvittaessa ontologiaa laajennetaan sovelluskohtaisilla luokilla, esimerkiksi jos halutaan mallintaa kysymyksiin vastaajia.

5.3.2 Tapausten haku

Kuvassa 5.2 tapausten hakuaskel on kuvattu vihreällä nuolella. Käyttäjä syöttää järjestelmään kysymyksen, ja se syötetään semanttiselle tiedoneristäjälle, joka palauttaa kysymyksestä löydetyt käsitteet, ilmentymät sekä henkilöiden ja paikkojen nimet. Kuvassa tästä käytetään termiä *käsitteiden louhinta*.

Asiasanalista järjestetään oletetun relevanssin mukaan. Kullekin asiasanaehdotukselle lasketaan painokerroin seuraavaa kaavaa käyttäen:

$$paino = tf * idf * klikkiKerroin * erisnimiKerroin \tag{5.1}$$

Kaavassa *paino* on asiasanan saama paino, tf käsitteen esiintymiskerrat kysymyksessä, *idf* käänteinen lukumäärä käsitteen käyttökerroista muissa kysymys-vastauspareissa ja $klikkiKerroin$ kerroin, joka ilmaisee kuuluuko asiasana johonkin semanttiseen klikkiin asiasanajoukossa. Näiden semanttisten klikkien päättelyyn käytetään sanasto-ontologiassa määriteltyjä semanttisia suhteita. Jos asiasana on paikka, ilmentymä tai henkilönnimi, kasvatetaan painoa antamalla tekijän $erisnimiKerroin$ arvoksi jokin yhtä suurempi luku. Perusteluna tässä on se, että erisnimet ja ilmentymät ovat tarkkuutensa takia todennäköisesti hyviä asiasanoja.

Varsinaisten kysymystekstistä löytyneiden käsitteiden lisäksi asiasanalistaan voidaan lisätä asiasanoja, joita on käytetty tapauskannassa olevissa kysymys-vastauspareissa usein asiasanoina kysymystekstistä löytyneiden asiasanojen kanssa.

Käyttäjän tehtävänä tässä askeleessa on valita semanttisen tiedoneristäjän ehdottamista asiasanoista kysymyksen kannalta olennaiset asiasanat. Näitä käyttäjän vahvistamia asiasanoja käytetään haettaessa tapauskannasta samankaltaisia kysymyksiä. Jos käyttäjä ei valitse yhtään asiasanaa, käytetään kaikkia asiasanaehdotuksia hakemiseen. Haku kohdistetaan kysymystekstiin sekä niiden asiasanoihin siten, että hakuosumat asiasanoihin ovat merkityksellisempiä kuin hakuosumat kysymystekstiin. Hakutulokset järjestetään relevanssinsa mukaan siten, että mitä enemmän asiasanoja niihin osuu, sitä suurempi relevanssi niillä on.

Tapausten haku -askeleen viimeinen vaihe on, että hakutulokset esitetään käyttäjälle relevanssijärjestyksessä.

5.3.3 Löydettyjen tapausten hyödyntäminen

Tapausten haku -askeleen lopputulemana on relevanssin mukaan järjestettyjä vanhoja kysymys-vastauspareja. Tässä askeleessa käyttäjän tehtävänä on päättää, haluaako hän käyttää jotain vanhaa kysymys-vastausparia vastaamisen pohjana. Kuvassa 5.2 on esitetty, miten käyttäjä on valinnut listalta yhden kysymys-vastausparin (kuvassa lihavoitu), ja sen vastaus on kopioitu uuden vastauksen pohjaksi.

Lopuksi käyttäjä muokkaa vastauksen haluamallaan tavalla tai laatii tyhjästä uuden, jos tapauskannasta ei löytynyt oleellisia vanhoja kysymys-vastauspareja.

5.3.4 Ratkaisun arviointi

Ratkaisun arviointi -askeleessa käyttäjän laatimaa vastausta rikastetaan resursseilla erilaisista ulkoisista tietolähteistä. Resurssien haku tapahtuu tapausten haku -askeleessa laaditun vahvistetun asiasanalistan perusteella, ja käyttäjän tehtävänä on valita, mitä resursseja hän haluaa käyttää vastauksessa. Ulkoiset aineistot voivat olla esimerkiksi sovellusalan kannalta mielenkiintoisia linkkikirjastoja, aineistotietokantoja tai sovellusalan yleistä tietämystä. Ulkoisten aineistojen hakutulokset ovat parempia, jos käytetyt tietolähteet ovat semanttisesti yhteensopivia mallin kanssa, esimerkiksi jos ne on annotoitu mallin käyttämällä sanasto-ontologialla tai -ontologioilla.

Käytännössä ratkaisun arviointi ja löydettyjen tapausten hyödyntäminen -askeleet voivat tapahtua samanaikaisesti. Tässä mallissa ne on kuitenkin eroteltu eri askeleisiin selkeyden vuoksi.

5.3.5 Ratkaisun säilöminen

Ratkaisun säilöminen -askeleessa kysymys-vastauspari tallennetaan tapauskantaan. Malli ei ota kantaa siihen, miten tämä tapahtuu. Askel voi sisältää esimerkiksi jonkin hakemiston päivittämisen tai kysymys-vastausparin liittämisen ulkoiseen aineistoon. Mahdollista on myös jonkinlainen *adaptiivisuus*, eli käyttäjän tekemien asiasana- ja resurssivalintojen perusteella tehdään päätelmiä käyttäjän profiilista, ja mallin muita askeleita mukautetaan oletetun profiilin mukaan.

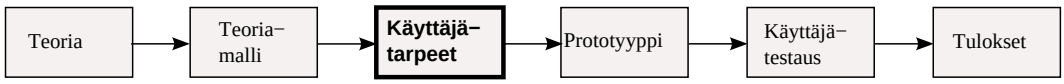
5.4 Mallin arviointia

Tässä kappaleessa esitetyssä mallissa semanttista asiasanoitusta hyödyntävästä kysymys-vastauspalvelusta on intuitiivisella tavalla yhdistetty tapauksiin perustuvan päättelyn teoriaa semanttisen webin tekniikoihin, erityisesti ontologioihin sekä semanttiseen asiasanoitukseen. Mallissa ihmisellä on luonteva rooli osana CBR-syklin askeleita.

Mallissa sovelletaan kuitenkin tapauksiin perustuvaa päättelyä varsin yleisellä tasolla, eikä eri askelten toteuttamiseen syvennytä tarkemmin. Malli ei ota kantaa siihen, miten sanasto-ontologiassa määriteltyjä semanttisia suhteita hyödynnetään esimerkiksi tapausten haku -askeleessa.

Luku 6

Käyttäjätarpeiden määrittely



Edellisessä kappaleessa luotiin teoreettinen malli tapauksiin perustuvaa päättelyä ja semanttista asiasanoitusta käyttävästä kysymys-vastauspalvelusta. Tässä kappaleessa on kuvattu käyttäjätarpeiden määrittely, jolla selvitettiin kirjastonhoitajien ja Kirjastot.fi-toimituksen tarpeita ja toiveita Kysy kirjastonhoitajalta -palvelun kehittämiseksi. Ensin esitellään olemassa oleva palvelu, minkä jälkeen esitellään vaatimuskartoitus ja sen tulokset.

6.1 Olemassa olevan palvelun kuvaus

Kysy kirjastonhoitajalta[1] on suomalainen Kirjastot.fi-toimituksen ylläpitämä palvelu, jossa kansalaiset voivat kysyä kirjastonhoitajilta kysymyksiä ja saada vastauksen kolmen arkipäivän sisällä. Palvelussa on mukana 39 kunnankirjastoa ja 14 muuta kirjastoa ja tietopalvelua. Näihin 14 erikoiskirjastoon kuuluuvat muun muassa Näkövammaisten kirjasto ja Eduskunnan kirjasto. Kysymykset lähetetään web-lomakkeen kautta kirjastoalan asiantuntijoille, jotka vastaavat niihin sähköpostitse. Kysymyslomakkeessa syötetään varsinaisen kysymyksen lisäksi myös asuinkunta, ja vastaamispäätös tehdään pääpiirteittäin sen mukaan: espoolaiset kirjastonhoitajat vastaavat espoolaisten ja sipoolaiset sipoolaisten asiakkaiden kysymyksiin.

[1] `http://www.kirjastot.fi/tietopalvelu`

Kirjastonhoitajien käyttämässä vastaamiskäyttöliittymässä on tekstikentät kysymys- ja vastaustekstiä varten. Varsinaisen vastauksen lisäksi kirjastonhoitajat liittävät kysymys-vastauspariin asiasanoja Yleisestä suomalaisesta asiasanastosta[2] (YSA) kuvaamaan kysymyksen ja vastauksen sisältöä. Kysymystä ja vastausta ei asiasanoiteta erikseen, vaan asiasanat kuvaavat koko kysymys-vastausparin sisältöä. Asiasanat kirjoitetaan tekstikenttään pilkulla eroteltuna. Lisäksi vastaukseen voi liittää niin sanotun *piilovastauksen*, joka lähetetään vastaussähköpostiin mukaan, mutta ei palvelun arkistoon. Näin kirjastonhoitaja voi liittää vastaukseen henkilökohtaisia huomautuksia, jotka eivät kuitenkaan suurta yleisöä kiinnosta. Vastatut kysymykset löytyvät arkistosta, johon voi suorittaa yksinkertaisia sanahakuja sekä päivämäärän mukaan rajattuja hakuja. Arkistossa on tällä hetkellä noin 20000 kysymys-vastaus-paria, ja uusia vastauksia lisätään jatkuvasti.

Seuraavassa on esimerkki yhdestä kysymyksestä ja vastauksesta palvelussa:

> **Kysymys:** Kalannahan muokkauksesta kertovaa kirjallisuutta. löytyykö?
>
> **Vastaus:** Kalannahan muokkauksesta löytyy suomeksi tietoa ainakin kirjasta: Eskelinen, Jouko: Harrastajanahkurin käsikirja 1. Myös muista nahankäsittelykirjoista voi löytyä tietoa aiheesta. Englanniksi kirjasta Churchill: The complete book of tanning skins and furs. Kalannahan käsittelystä löytyy myös joitakin lehtiartikkeleita, joita voit kysyä kirjastosta. Esim. Taito-lehden numero 1 v. 1997, sivut 40-41: Kalannahkaa parkitsemaan: lohi, kuha tai made.
>
> **Asiasanat:** kalat nahka nahkatyöt parkitus

Jotkut palveluun tulevat kysymykset ovat yksinkertaisia ja kirjastonhoitaja voi vastata niihin välittömästi. Nämä kysymykset käsittelevät muun muassa kirjastojen aukioloaikoja tai tiedusteluja kaukolainojen tekemisestä. Useimmat kysymykset vaativat kuitenkin enemmän panostusta vastaamistyöhön. Tällaisia kysymyksiä ovat esimerkiksi yllä esitetyn kaltaiset kysymykset, joissa vastausta varten voidaan tehdä selviys useastakin eri lähteestä ja vastauksen pituus voi olla useita kappaleita pitkä.

Kysymys-vastauspariin liitettävät asiasanat kertovat, mihin aihepiiriin kysymys-vastauspari liittyy, ja niiden avulla voi selata palvelussa olevia muita kysymyksiä. Käyttäjä voi esimerkiksi *kalat* asiasanaa klikkaamalla nähdä

[2] `http://vesa.lib.helsinki.fi`

muut kysymys-vastausparit, joihin kyseinen asiasana liittyy. YSA-asiasanojen lisäksi asiasanoina käytetään usein henkilöiden nimiä.

Arkistoon on yksinkertainen sanahaku, joka on esimerkiksi Googlen[3] vastaavaa kehittyneempi siinä mielessä, että siinä voi käyttää jokerimerkkejä $*$ ja ? ($*$ vastaa mitä tahansa merkkijonoa ja ? mitä tahansa merkkiä). Sanahakua ei voi kuitenkaan kohdistaa esimerkiksi asiasanakenttään eikä siinä ole kehittyneempiä boolen hakuoperaattoreita.

6.2 Tiedonkeruumenetelmät

Käyttäjätarpeiden määrittelyä varten tietoa kerättiin seuraavilla tavoilla:

- Keskustelut Kirjastot.fi-toimituksen edustajien kanssa.
- Haastattelut Kysy kirjastonhoitajalta -palvelua käyttävien kirjastonhoitajien kanssa.
- Kysy kirjastonhoitajalta -palvelun avulla lähetetty kysely koko palvelun kirjastonhoitajille.

Kirjastot.fi-toimituksen edustajien kanssa käytyjen keskustelujen perusteella määriteltiin suurpiirteiset linjaukset siitä, mihin suuntaan Kysy kirjastonhoitajalta -palvelua haluttiin kehittää. Tämän lisäksi käyttäjien tarpeiden kartoittamiseksi haastateltiin viittä Kysy kirjastonhoitajalta -palvelua käyttävää kirjastonhoitajaa ja lisäksi tehtiin yksi sähköpostikysely. Haastattelujen sisältö käsitteli Kysy kirjastonhoitajalta -palvelua yleisesti, sekä erityisesti asiasanoitusta. Kussakin haastattelussa oli läsnä yksi haastateltava, paitsi yhdessä kaksi haastateltavaa.

Vapaat haastattelut ovat yleensä sopivimpia, kun tavoitteena on saada yleistä tietoa käyttäjistä, ja kun käyttäjien ongelmista ei ole tarkempaa tietoa [Xri00]. Koska tarkoituksena oli kuitenkin selvittää tiettyjä asioita, kuten asiasanoituksen helppoutta ja tiedonhaun ongelmakohtia, valittiin haastattelumenetelmäksi puolistrukturoitu haastattelu, jossa keskustelua ohjaa haastattelurunko, mutta aiheet voivat elää runsaastikin haastateltavan mielenkiinnon mukaan. Haastattelun runkona käytetty kysymyslista löytyy liitteestä 1.

Haastatellut henkilöt edustivat kolmea käyttäjäryhmää:

[3] `http://www.google.fi`

- Neljä kirjastonhoitajaa, jotka edustivat tyypillisiä kysymyksiin vastajia. (kolme naista, yksi mies)
- Valtakunnallisella tasolla toimiva kysymyksiin vastaaja, jolla oli pitkä kokemus palvelusta. (nainen)
- Palvelun ylläpitäjä ja suunnittelija Kirjastot.fi-toimituksessa (nainen)

Haastattelemalla erilaisia käyttäjiä pyrittiin saamaan mahdollisimman monipuolinen näkemys palvelusta ja sen kehittämiskohteista. On syytä kuitenkin huomata, ettei työn laajuuden puitteissa ollut mahdollista tehdä kattavaa käyttäjätarve- ja vaatimusmäärittelyä. Esimerkiksi palvelun loppukäyttäjiä ei haastateltu ollenkaan, joskin kirjastonhoitajien kanssa käydyissä keskusteluissa tuli esiin myös loppukäyttäjien käyttämä hakutoiminto.

Käyttäjähaastattelujen lisäksi lähetettiin kysely Kysy kirjastonhoitajalta -palvelun kautta. Kyselyn tarkoituksena oli tavoittaa käyttäjäkuntaa laajemmalti. Kysymys kuului: *Miten palvelun arkistosta hakua kannattaisi käyttää mahdollisimman tehokkaasti? Miten kirjastonhoitajat haluaisivat parantaa hakutoimintoa, eli millaisia puutteita arkistohaussa on?*

6.3 Tulokset

Seuraavassa esitellään käyttäjätarpeiden määrittelyn tuloksia tiedonkeruumenetelmien mukaan jaoteltuna.

6.3.1 Keskustelut Kirjastot.fi-toimituksen edustajien kanssa

Työn alussa käytyjen keskustelujen perusteella Kirjastot.fi-toimituksen tärkeimmät tavoitteet kehitystyön suhteen olivat:

1. Helpottaa kirjastonhoitajien vastaamistyötä.
2. Parantaa eri tietovarastojen integrointia.
3. Tehostaa hakutoimintoja.

Kirjastonhoitajan vastaamistyön helpottamisella tarkoitettiin sitä, että vastausten luominen ei vaatisi niin paljon vaivaa. Erityisesti vastausten asiasanoittaminen koettiin hankalaksi. Tietovarastojen integroinnin parantamisella tarkoitettiin sitä, että Kirjastot.fi-toimituksen käyttämät tietolähteet, kuten esimeriksi Kysy kirjastonhoitajalta, Linkkikirjasto[4] ja Tiedonhaun portti[5] toimisivat paremmin yhteistyössä. Hakutoimintojen tehostamista ei tarkemmin määritelty, vaan sillä tarkoitettiin yleisesti saannin ja tarkkuuden parantamista. Tavoitteista erityisesti vastaamistyön helpottamista pidettiin tärkeänä kehityskohteena.

6.3.2 Kysely Kysy kirjastonhoitajalta -palveluun

Palveluun lähetettyyn kysymykseen saatiin kaksi vastausta. Toisen vastauksen oli laatinut kaksi kirjastonhoitajaa, toisen yksi. Vastaukset tulivat pienehköistä Suomen kaupungeista.

Molemmissa vastauksissa ehdotettiin, että olisi hyvä, jos palvelu tarkastaisi automaattisesti, onko tulevaan kysymykseen vastattu jo aikaisemmin. Toisessa vastauksessa ehdotettiin, että arkistohaussa pitäisi olla kehittyneempiä hakumahdollisuuksia, etenkin haun kohdistaminen asiasanakenttään. Molemmissa vastauksissa kritisoitiin hieman nykyisiä hakutoiminnallisuuksia: vastaajat kyseenalaistivat *Hae viimeisen päivän aikana vastatut* tai *Hae aikavälillä* -toiminnallisuuksien tarpeellisuuden.

6.3.3 Haastattelut

Seuraavassa on eritelty alakohdittain haastatteluissa esiintyneitä yhteisiä piirteitä ja mielipiteitä.

Palvelun hyviä puolia

Lähes kaikki vastaajat pitivät Kysy kirjastonhoitajalta -palvelun hyvänä puolena sitä, että palvelussa on mukana niin monta vastaajakirjastoa, ja että palvelu kattaa koko maan kirjastot. Hyvänä puolena pidettiin myös sitä, että kirjastonhoitajalla on aikaa keskittyä vastaamiseen ja laatia vastauksia rauhassa useiden lähteiden perusteella.

[4] `http://www.kirjastot.fi/linkkikirjasto/`
[5] `http://tiedonhaunportti.kirjastot.fi`

Palveluun vastaamisen kerrottiin myös tukevan ammatillista itsetuntoa ja osaamista, kun hankaliin kysymyksiin saa laadittua hyvän vastauksen. Mielenkiintoinen tähän liittyvä piirre oli se, että eräs vastaaja kertoi miettivänsä vastaamisessa myös sitä, miltä vastaus näyttää muiden kirjastonhoitajien silmissä.

Kirjastonhoitajien käytössä on myös kaikki vastaajat sisältävä sähköpostilista, jolle voi lähettää tiedusteluja vaikeista kysymyksistä. Tämän listan käytön koettiin kasvattavan kollektiivista ammattiosaamista, esimerkiksi yksi haastatelluista oli vaikuttunut, miten nopeasti joskus tulee vastaus vaikeaan runouteen liittyvään kysymykseen.

Palvelun huonoja puolia

Siinä missä rauhassa vastaaminen koettiin hyväksi puoleksi, niin lähes kaikki haastatellut mainitsivat kuitenkin palvelun huonoksi puoleksi sen, että kontakti asiakkaaseen puuttuu. Tästä suoran asiakaskontaktin puutteesta seuraa se, ettei asiakkaalta voida kysyä nopeasti tarkentavia kysymyksiä. Lisäksi usein kasvokkain kommunikoidessa jo pelkästä äänenpainosta ja ruumiinkielestä voi päätellä, mitä asiakas haluaa ja tarkoittaa kysymyksellään. Usein kysymykseen esitetäänkin jatkokysymyksiä sähköpostitse, mutta tämä sähköpostikirjeenvaihtoa ei säilötä mihinkään eikä sitä liitetä vastaukseen. Voi myös käydä niin, että epätäydellinen vastaus unohtuu suljettuun arkistoon, vaikka vastaus sähköpostikeskustelujen perusteella onkin sittemmin täydentynyt.

Vastaamisesta

Päätös kysymykseen vastaamisesta riippuu siitä, millaisessa roolissa kirjastonhoitaja on palvelussa. Esimerkiksi erikoiskirjaston vastaaja kertoi yksinkertaisesti vastaavansa kysymyksiin, jotka hänen organisaatiolleen on osoitettu. Palvelun valtakunnallinen ylläpitäjä taas vastaa kysymyksiin, joihin muut eivät vastaa. Loput, ”tavalliset” kirjastonhoitajat vastaavat kunnalle osoitettuihin kysymyksiin mielenkiinnon ja osaamisen mukaan.

Vastaamiseen käytetty aika vaihtelee suuresti. Eräs vastaaja kertoi, että joskus voi käyttää kokonaisen työpäivänkin yhteen kysymykseen vastaamiseen. Toisinaan vastauksen osaa laatia heti, etenkin jos sen tietää heti tutustumatta lähdekirjallisuuteen. Vaikeimmiksi kysymyksiksi vastata mainittiin runouteen liittyvät kysymykset.

Kirjastonhoitajat käyttävät monipuolisesti lähteitä vastaamiseen, kuten käsikirjastoa, Helmet-asiasanahakua, Frank-monihakua, eri kirjastojen aineistotietokantoja, palvelun arkistoa sekä yleisiä hakukoneita, kuten Googlea.

Asiasanoituksesta

Asiasanoituksessa on hyvin erilaisia käytäntöjä haastateltujen kirjastonhoitajien välillä. Yksi haastateltu ei käytä asiasanoja ollenkaan, kun taas yksi valitsee ensin asiasanojen kautta näkökulman vastaamiseen, ja laatii vastauksen vasta sen jälkeen. Helpoksi asiasanoittamisessa koettiin nimet (esimerkiksi kysymykset liittyen Aleksis Kiveen) ja vaikeaksi ylä-alakäsitteisiin liittyvä problematiikka (valitako asiasanaksi Aleksis Kivi vai kirjailijat).

Useimmat haastatellut pitivät asiasanoitusta tärkeänä ominaisuutena, koska sen perusteella tehdään hakuja arkistoon. Yksi haastatelluista muistutti kuitenkin, että asiasanoitusta ei pitäisi nähdä itseisarvona vaan pitäisi miettiä, miksi sitä ylipäänsä tehdään.

Ominaisuusideoita

Haastattelun yhteydessä kysyttiin myös, mitä uusia ominaisuuksia palveluun kirjastonhoitajat haluaisivat, jos voisivat vapaasti päättää. Seuraavassa on luokiteltu näitä ominaisuusideoita sen mukaan, mihin toiminnallisuuteen ne liittyvät:

Yleisiä palveluun liittyviä toiveita

- Palvelun ulkonäkö on siisti, mutta tavallinen. Voisi olla jollain tapaa markkinointihenkisempi ja hienompi.
- Palvelussa pitäisi olla enemmän tietoa eri tavoista kysyä kirjastonhoitajilta.
- Palvelussa voisi olla jokin keino tavoittaa kysyjä nopeammin kuin sähköpostilla. Joskus sähköpostiin vastaaminen voi viedä viikonkin asiakkaalta.

Kysymyksiin vastaaminen

- Kysymykselle voisi olla uusi tila: “vastattu, mutta ei vastattu täydellisesti”.

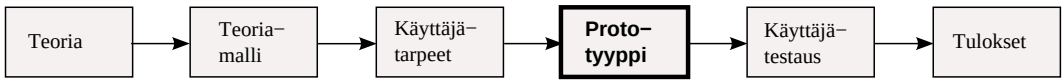
- Kirjastonhoitajalla kirjastossa mahdollisuus lisätä nopeasti kysymys ja vastaus palvelun arkistoon, kun asiakas tulee kysymään kirjastossa kasvokkain.
- Sähköpostitse käydyn kommunkoinnin liittäminen kysymys-vastauspariin.
- Jonkinlainen huomautus tai ilmoitus muille vastaajille, jos joku kirjastonhoitaja varaa kysymyksen, mutta sitten jostain syystä ei vastaakaan kysymykseen.
- Valittavat asiasanat voisivat olla kokonaisina sanoina pudotusvalikossa, jolloin niitä voisi valita useammankin kerralla.
- Asiasanoituksessa asiasanoja pitäisi voida yhdistää JA-relaatiolla.
- Toiminnallisuus, joka tarkastaa, onko tulevaan kysymykseen vastattu jo aiemmin.

Arkisto ja hakutoiminto

- Arkiston roolia voisi miettiä; onko kyseessä arkisto vai tietämyspankki? Esimerkiksi viisi vuotta sitten vastatut kysymykset eivät välttämättä ole tänä päivänä valideja.
- Arkistossa on vanhentuneita kysymyksiä. Esimerkiksi vuonna 1999 kysyttyjä kysymyksiä voisi jo poistaakin.
- Arkistohaussa voisi olla kirjastonhoitajalle enemmän yhdistelymahdollisuuksia, kuten JA- ja TAI-operaattoreita. Asiakkaalle haku olisi hyvä säilyttää yksinkertaisena.
- Haun kohdistaminen asiasanakenttään.

Luku 7

Opas – Kysy kirjastonhoitajalta semanttisessa webissä



Edellisessä kappaleessa esitettiin työtä varten tehty käyttäjätarpeiden määrittely. Tässä kappaleessa kuvataan näihin tarpeisiin, kappaleessa 5 esitettyyn teoreettiseen synteesiin sekä Kysy kirjastonhoitajalta -palveluun perustuva prototyyppi kysymys-vastauspalvelusta semanttisessa webissä, *Opas* [VHA06, VAH06].

7.1 Lähtökohdat Oppaalle

Käyttäjätarpeiden määrittelyssä tuli ilmi, että yleisesti Kysy kirjastonhoitajalta -palveluun oltiin tyytyväisiä, etenkin sen valtakunnallisuuden ja toimivuuden vuoksi. Asiasanoitus koettiin yleensä tärkeäksi, ja kirjastonhoitajat tuntuivat käyttävän palvelun hakutoimintoa erityisesti asiasanoja käyttäen. Useat haastatellut olivat sitä mieltä, että kysymykset palvelussa toistuvat, joten jonkinlainen samankaltaisten kysymysten etsijä on tarpeellinen. Vastaajat myös kertoivat käyttävänsä monipuolisesti lähteitä vastaamisen tukena, mistä voitiin päätellä, että eri tietolähteiden yhdistämisestä vastaustyökaluun voi olla hyötyä.

Näistä lähtökohdista Oppaaseen päätettiin toteuttaa puoliautomaattinen asia-

sanoittaja, joka auttaa kirjastonhoitajaa asiasanojen valinnassa. Lisäksi toteutettiin komponentti, joka automaattisesti etsii asiasanojen perusteella vanhoja samankaltaisia kysymyksiä ja muita mielenkiintoisia resursseja.

7.2 Oppaan yleisrakenne ja teknologiat

Oppaan rakenne perustuu yhteiseen Semantic Computing - tutkimusryhmässä[1] kehitettyyn Java-pohjaiseen ohjelmistoarkkitehtuuriin, joka tarjoaa tutkimusryhmässä tehtäville sovelluksille yhteisen sovelluskehyksen. Ohjelmistoarkkitehtuuri on esitetty kuvassa 7.1. Yhteinen komponenttipohjainen sovelluskehys hyödyttää sitä käyttäviä sovelluksia siten, että yksinkertaisia, sovelluksen kannalta epäolennaisia toiminnallisuuksia ei tarvitse toteuttaa, vaan ne tulevat sovelluskehyksen tarjoamana. Esimerkiksi jos jotain Oppaan käyttämistä ontologioista muokataan, huomaa sovelluskehys muuttuneen ontologian automaattisesti, ja lataa ontologiamallin uudelleen muistiin.

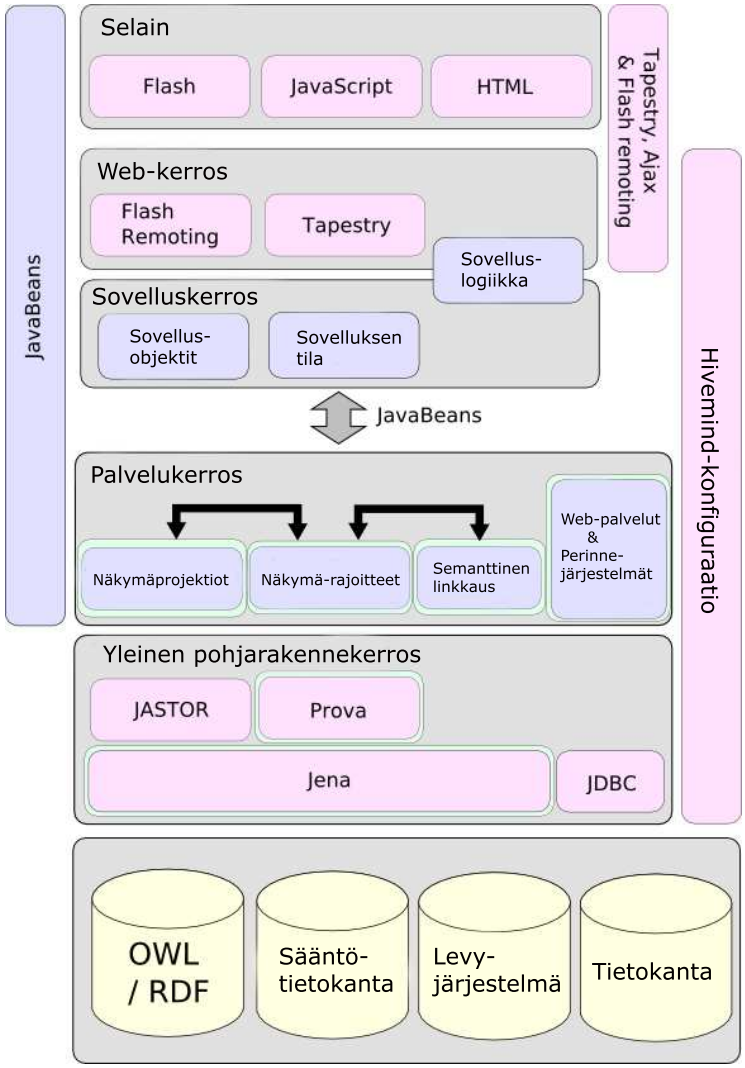
Taulukko 7.1: Oppaan käyttämät ulkoiset kirjastot ja ohjelmistot

Teknologia	**Käyttökuvaus**	**Lisätietoa**
MySQL	Ainestojen talletukseen käytetty tietokanta	http://www.mysql.com
Tapestry	Ohjelmistoarkkitehtuurin pohja, näkymät ja navigaatio	http://tapestry.apache.org/
Tacos	Tapestryn kanssa yhteensopivat Ajax-toiminnallisuudet	http://tacos.sourceforge.net
Jena	Ontologiamallien käsittely ja tallennus	http://jena.sourceforge.net
Jastor	Java-luokkarakenteen luonti ontologiamallien pohjalta	http://jastor.sourceforge.net
Prototype	Javascript-toiminnallisuuksien tekeminen	http://prototype.conio.net

Taulukossa 7.1 on esitetty Oppaan käyttämät ulkoiset kirjastot ja ohjelmistot. Oppaan toiminnallisuuksien toteuttamiseen on käytetty runsaasti myös niin sanottua Asynchronous JavaScript and XML[2] (Ajax) -tekniikkaa Perinteiset

[1] `http://www.seco.tkk.fi/`

[2] `http://en.wikipedia.org/wiki/AJAX`



Kuva 7.1: Oppaan pohjana toimiva ohjelmistoarkkitehtuuri. (Kuva: Eetu Mäkelä)

web-sovellukset perustuvat sivupyyntöihin ja -vastauksiin. Kun käyttäjä tekee jotain käyttöliittymässä, tehdään uusi pyyntö ja sivu ladataan uudelleen. Yksinkertaistaen sanottuna Ajaxin avulla yhteys palvelimelle pidetään auki yhden sivupyynnön aikana, jolloin käyttäjän tekemien käyttöliittymävalintojen perusteella voidaan palvelimelta hakea uutta tietoa ja muokata sivun XML-rakennetta. Tämä sallii yleensä vain työpöytäsovelluksista tuttujen monipuolisten käyttöliittymätoiminnallisuuksien toteuttamisen.

7.2.1 Aineiston ontologisointi

Sovellusta varten luotiin yksinkertainen ontologia kappaleessa 5 esitetyn mallin perusteella. Tämä ontologia kuvastaa palvelussa olevia kysymyksiä, vastauksia sekä asiasanoja. Kysy kirjastonhoitajalta -palvelun aineisto saatiin XML[3]-tiedostoina, joista muunnos tehtiin XSLT[4]-muunnoksia käyttäen.

Mahdollisuus tehdä koneellista päättelyä eksplisiittisesti määriteltyjen semanttisten suhteiden avulla on yksi ontologioiden eduista, kuten kappaleessa 2.2.1 todettiin. Oppaassa ei kuitenkaan kysymys-vastausontologian perusteella tehty minkäänlaista koneellista päättelyä. Tästä syystä erillistä ontologiaa ei olisi välttämättä tarvinnut tehdä, vaan aineiston olisi voinut syöttää relaatiotietokantaan ja käsitellä sitä suoraan, tallentamatta erilliseen ontologiamalliin. Ontologian käyttöön kuitenkin päädyttiin pääasiassa sen takia, että se tarjoaa hyvän rajapinnan sovelluksen ja sovelluksen käyttämän aineiston välille. Opas ei ota minkäänlaista kantaa siihen, minkälaista kysymys-vastausdataa se käyttää – riittää, että kysymys-vastausdata on muunnettu Oppaan käyttämän kysymys-vastausontologian mukaiseksi. Ontologiamallin käyttämisestä seuraa myös se etu, että sen ja Jastorin avulla voidaan automaattisesti luoda Java-luokkia, joita sovelluksen on helppo käyttää. Tämä ei sinänsä ole argumentti ontologiavalinnan puolesta, koska myös relaatiotietokannoille on olemassa runsaasti kehyksiä[5], jotka hoitavat peilauksen relaatiotauluista Java-olioiksi.

7.2.2 Ulkopuolisten aineistojen ontologisointi ja integrointi

Varsinaisen kysymys-vastausaineiston lisäksi sovellukseen tuotiin ulkopuolisia aineistoja:

Helsingin kirjaston luokitusjärjestelmä[6] (HKLJ) , jossa on määritelty, miten aineisto on luokiteltu Helsingin kirjastoissa. HKLJ:ssä on myös kuvattu mikä luokka liittyy mihinkin asiasanaan. Esimerkiksi asiasana *kokonaisluvut* liittyy kirjastoluokkaan *512 Lukuteoria. Aritmetiikka.* Tämän lisäksi asiasanoille on määritelty näkökulmia. Esimerkiksi asiasanaa *koivu* voidaan tarkastella muun muassa metsänhoidon tai puutöiden näkökulmasta. Kukin näkökulma liittyy myös johonkin kirjastoluokkaan.

[3] `http://www.w3.org/XML/`
[4] `http://www.w3.org/TR/xslt`
[5] Esimerkiksi Hibernate, `http://www.hibernate.org/`

Linkkikirjasto Linkkikirjastoon[7] on koottu erilaisia linkkejä, jotka on luokiteltu yleisten kirjastojen luokittelujärjestelmän (YKL)[8] perusteella.

Aineistot muunnettiin XML-muodosta RDF-muotoon kappaleessa 2.2.3 esiteltyjen periaatteiden mukaan [vAMS+04].

7.3 Oppaan toiminnallisuudet

Oppaaseen toteutettiin kaksi kirjastonhoitajan pääkäyttötapausta: vastaamattomien kysymyksien selailu ja kysymykseen vastaaminen. Oppaasta jätettiin toteuttamatta tutkimuskysymysten kannalta epäolennaisia toiminnallisuuksia, kuten sähköpostin lähettäminen kysyjille, olemassaolevien vastausten muokkaaminen ja niin edelleen.

7.3.1 Asiasanoitus

Oppaan yksi päätoiminnallisuuksista on puoliautomaattinen asiasanoitus. Tähän Opas käyttää tutkimusryhmässä kehitettyä *Pokaa*[9], joka on työväline semanttisen annotointiin. Pokalle annetaan syötteenä tekstinpätkä, ja se palauttaa vastauksena listan ontologisia käsitteitä, jotka se on löytänyt syötetekstistä. Oppaan tapauksessa Poka käyttää käsitteiden etsimisen perustana Yleistä suomalaista ontologiaa[10] [HVK+05] (YSO), joka on Yleisen suomalaisen asiasanaston[11] semanttinen versio. Minkä tahansa muun ontologian käyttäminen on mahdollista, mikä voisi tulla kyseeseen esimerkiksi, jos sovellusta muunnettaisiin muille kielille. YSO:sta löytyvien substantiivikäsitteiden lisäksi Poka tunnistaa syötetekstistä paikkoja ja henkilöiden nimiä. Paikkojen tunnistaminen perustuu erilliseen paikkaontologiaan[12]. Pokan palauttamat henkilönnimet ovat käytännössä YSO:n *henkilöt* -luokan ilmentymiä.

[7] http://www.kirjastot.fi/linkkikirjasto
[8] http://ykl.kirjastot.fi/
[9] http://www.seco.tkk.fi/applications/poka/
[10] http://www.seco.tkk.fi/ontologies/yso/
[11] http://vesa.lib.helsinki.fi/ysa/
[12] `http://www.seco.tkk.fi/ontologies/`

Asiasanaehdotusten järjestäminen

Asiasanaehdotusten järjestäminen tehtiin kappaleessa 5 esitetyn mallin perusteella. Semanttisten klikkien päättelyyn käytettiin YSO:ssa määriteltyjä ylä- ja alakäsite -suhteita (*koira* on *eläimen* alakäsite) sekä niin sanottuja lähikäsitesuhteita (käsite *jalankulkijat* liittyy käsitteeseen *kevyt liikenne*)

Seuraavassa on tarkemmin eritelty, miten kukin näistä kertoimista laskettiin:

Asiasanan esiintymiskerrat tekstissä, tf

Jos asiasana löytyy tekstistä useasti, on syytä olettaa sen olevan olennainen tekstiin nähden. Alustavien testien perusteella vaikutti siltä, että asiasanan esiintymiskerroilla tulisi olla suuri painoarvo painokertoimen laskemisessa. Näin ollen tf laskettiin kaavalla

$$tf = f^{1.6} \tag{7.1}$$

missä f on asiasanan esiintymiskerrat tekstissä. Näin asiasanan painokerroin kasvaa eksponentiaalisesti esiintymiskertojen suhteen. Tällä saavutettiin se, että jos asiasana esiintyy useasti tekstissä, se varmasti saa suuren kertoimen, vaikka sitä oltaisiinkin käytetty usein asiasanana muissa kysymyksissä.

Käänteinen käyttökertojen lukumäärä, idf

Voidaan olettaa, että jos asiasanaa ei ole käytetty usein muissa kysymyksissä asiasanana, se on hyvä asiasana tiedonhaun kannalta. Näin ollen niiden asiasanojen, joita on käytetty usein, painoa pienennettiin käyttämällä kaavaa

$$idf = \frac{1}{0.2 * n} \tag{7.2}$$

missä n on asiasanan käyttökertojen lukumäärä muissa kysymyksissä. Kerroin 0.2 laitettiin nimittäjään, koska alustavien testien mukaan käyttökertojen määrän ei pitäisi vaikuttaa kovin paljon asiasanan painoon.

Semanttiset klikit, $klikkiKerroin$

Klikkikerrointa kerrottiin luvulla 2.5 kutakin muista asiasanoista löytynyttä liittyvää käsitettä kohden ja luvulla 1.5 kutakin löytynyttä ylä- tai alakäsitettä kohden. Esimerkiksi jos tekstistä löydetään asiasanat *hivenaineet*, *kivennäisaineet* ja *alkuaineet*, saa asiasana *hivenaineet* klikkikertoimekseen 3.75 ($1.5 * 2.5$), koska se on käsitteen *kivennäisaineet* alakäsite ja *alkuaineet* on siiten liittyvä käsite.

Erisnimet ja ilmentymät, *erisnimiKerroin*
Paikoille, henkilönnimille ja ilmentymille annettiin *erisnimiKerroin* 5, jotta niillä olisi suuri painoarvo samankaltaisia kysymyksiä ja muita resursseja etsittäessä.

Asiasanojen tarkentaminen

Kun kielenkäyttö kehittyy, syntyy sanastoon uusia käsitteitä, joita ei vielä virallisesti ole lisätty esimerkiksi YSO:on tai Yleiseen suomalaiseen asiasanastoon, mutta jotka asiasanoittamisen kannalta ovat olennaisia ja hyväksyttäviä. Tästä syystä Oppaassa on mahdollista *tarkentaa* asiasanoja. Tarkennukset voivat olla joko YSO:n käsitteiden *alakäsitteitä* (lintuinfluenssa on influenssan alakäsite) tai *ilmentymiä* (Lassie on koiran ilmentymä). Tarkentamalla luodut uudet käsitteet ja ilmentymät ovat myös muiden kirjastonhoitajien käytettävissä asiasanoituksessa.

Vapaat asiasanat

Usein käy niin, että asiasanaehdotusten joukossa ei ole asiasanoja, joita kirjastonhoitaja haluaa käyttää. Tästä syystä Oppaassa on mahdollista liittää kysymyksiin ja vastauksiin *vapaita asiasanoja*. Nämä asiasanat eivät ole täysin vapaita siinä mielessä, että asiasanojen tulee löytyä YSO:sta. Jos kirjastonhoitaja kuitenkin haluaa käyttää asiasanaa, jota ei löydy YSO:sta, hänen tulee aiemmin kuvatulla tavalla tarkentaa jotain YSO:sta löytyvää käsitettä.

Opaassa hyödynnettiin automaattista semanttista täydennystä [HM06] vapaiden asiasanojen lisäämiseen. Tämän tarkoituksena on varmistaa, että asiasanoituksessa käytetään ontologiassa olevia käsitteitä sekä myös ehdottaa kirjastonhoitajalle lisättävään asiasanaan liittyviä muita asiasanoja.

7.3.2 Vastaajan apurit

Vastaajien apureiksi kutsutaan Oppaaseen toteutettuja samankaltaisten kysymysten ja muiden resurssien etsimiskomponentteja. Vastaajan apureiden tarkoitus on auttaa kirjastonhoitajaa vastauksen laatimisessa yhdistämällä eri tietolähteitä Oppaaseen. Apureihin kuuluu kysymys-vastausarkisto, HKLJ-luokitusjärjestelmä sekä Linkkikirjasto. Yhteistä näille apureille on se, että niiden sisältö on riippuvainen Pokan ehdottamista asiasanoista, kirjastonhoi-

tajan tekemistä asiasanavalinnoista ja lisätyistä vapaista asiasanoista. Kustakin apurista voi liittää automaattisesti tekstiä vastaukseen.

7.4 Esimerkki prototyypin käytöstä

Kuvassa 7.2 on esitetty prosessimalli, jonka mukaan kirjastonhoitajan on tarkoitus vastata kysymyksiin Oppaassa. Seuraavassa tätä vastaamismallia havainnollistetaan näyttämällä, miten kysymykseen vastaataan Oppaan avulla. Siinä kirjastonhoitaja vastaa kysymykseen liittyen varusmiesten alkoholinkäyttöön. Prosessin ensimmäistä ja viimeistä askelta lukuunottamatta vaiheita ei ole pakotettu, joten kirjastonhoitaja voi laatia vastauksen ennen kuin valitsee asiasanat kysymykselle tai korjata kysymyksen kieliasun viimeisenä. Tämä ei kuitenkaan ole suositeltavaa, koska esimerkiksi vastaajan apureiden sisältö riippuu asiasanavalinnoista ja asiasanoja löydetään paremmin, jos kysymystekstissä mahdollisesti olevat kirjoitusvirheet on korjattu.

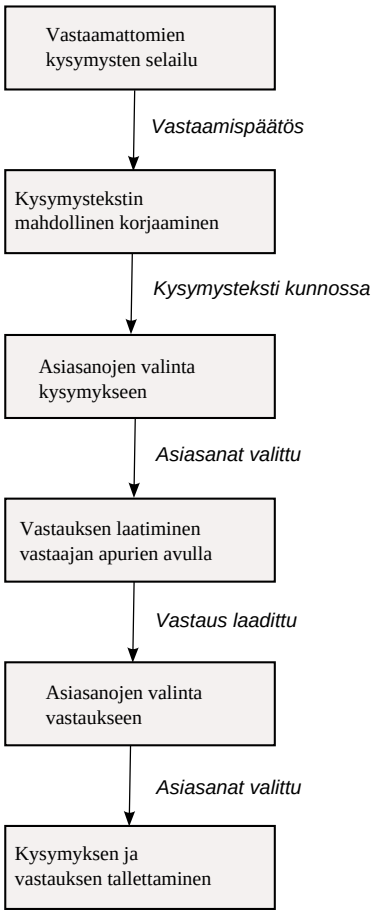
7.4.1 Kysymyksen valitseminen ja vastaamisnäkymä

Kuvassa 7.3 on havainnollistettu vastaamisprosessin ensimmäinen vaihe, vastaamattomien kysymyksien selailu. Tämän näkymän kautta kirjastonhoitaja voi selailla järjestelmään tulleita kysymyksiä, ja tehdä vastaamispäätöksen. *Käsittele*-linkkiä painamalla päädytään näkymään, joka on havainnollistettu kuvassa 7.4.

Kuvassa vasemmalla nähdään kysymysteksti, jota kirjastonhoitaja voi muokata. Kysymystekstilaatikon alla on linkki, jonka avulla voidaan etsiä uudelleen asiasanaehdotuksia siinä tapauksessa, että kysymystekstiä pitää muokata, esimerkiksi poistaa kirjoitusvirheitä. Näkymässä oikealla näkyvät asiasanaehdotukset, sekä linkki vapaiden asiasanojen lisäämiseksi. Kysymystekstin alla on vastaajan apurit.

7.4.2 Asiasanojen valinta

Kuvassa 7.5 on havainnollistettu asiasanaehdotukset sekä niiden valinta. Poka on löytänyt kysymyksestä kolme käsitettä (*muistot*, *alkoholinkäyttö* ja *vapaa-aika*). Suluissa on näytetty kunkin käsitteen yläkäsite. Tämän lisäksi Poka on löytänyt kaksi erisnimeä: henkilönnimen *Martti Ahtisaari* sekä paikan *Santa-*

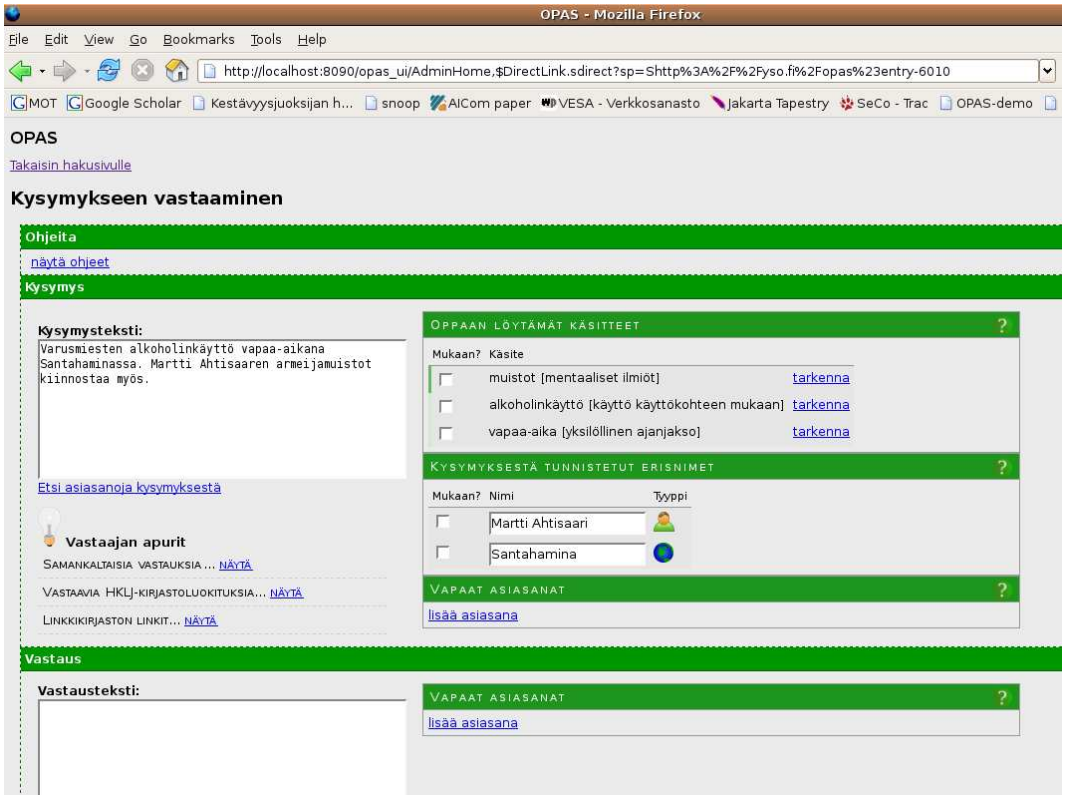


Kuva 7.2: Suunniteltu vastaamisprosessi Oppaassa

hamina. Kunkin käsitteen edessä on asetusnappi, jonka avulla kirjastonhoitaja voi ottaa käsitteen mukaan asiasanoitukseen. Kirjastonhoitaja on päätynyt liittämään kaikki asiasanat paitsi *muistot* kysymykseen.

Poka ei kuitenkaan ole löytänyt tekstistä käsitettä *varusmiehet*, joka on kysymyksen kannalta oleellinen. Kuvassa 7.5 kirjastonhoitaja on kirjoittamassa sanaa vapaan asiasanoituksen kenttään, ja samaan aikaan tekstikentän alla päivittyy lista vastaavista käsitteistä YSO:ssa. Kun kirjastonhoitaja vie kursorin täydentäjän ehdottaman käsitteen päälle, näytetään käsitteeseen liittyviä muita käsitteitä sekä ylä- ja alakäsitteet. Klikkaamalla käsitettä se voidaan liittää tekstikentään asiasanaksi. Tekstin punainen väri ilmaisee sen, että tekstikentässä sillä hetkellä oleva sana ei vastaa YSO:n käsitettä. Musta väri taas ilmaisee, että tekstikentässä oleva sana vastaa jotain YSO:n käsitettä.

Kuva 7.3: Vastaamattomien kysymyksien selailunäkymä Oppaassa



Kuva 7.4: Kysymykseen vastaaminen Oppaassa

Asiasanaehdotusten painoarvoja on visualisoitu siten, että mitä suurempi paino asiasanalla on, sitä voimakkaamman vihreä pieni palkki sen eteen laitetaan. Kuvassa 7.5 nähdään, miten asiasanalla *muistot* on suurempi paino kuin muilla

Kuva 7.5: Asiasanojen valinta Oppaassa

kahdella asiasanalla. Tämä johtuu siitä, että asiasanaa *muistot* ei ole käytetty usein asiasanana aikaisemmissa kysymyksissä.

Asiasanan tarkennus

Kirjastonhoitaja huomaa vielä, että kysymys käsittelee Martti Ahtisaaren armeijamuistoja. YSO:ssa ei ole käsitettä *armeijamuistot* ja niinpä kirjastonhoitaja tarkentaa asiasanaa *muistot* kuvassa 7.6 esitetyllä tavalla.

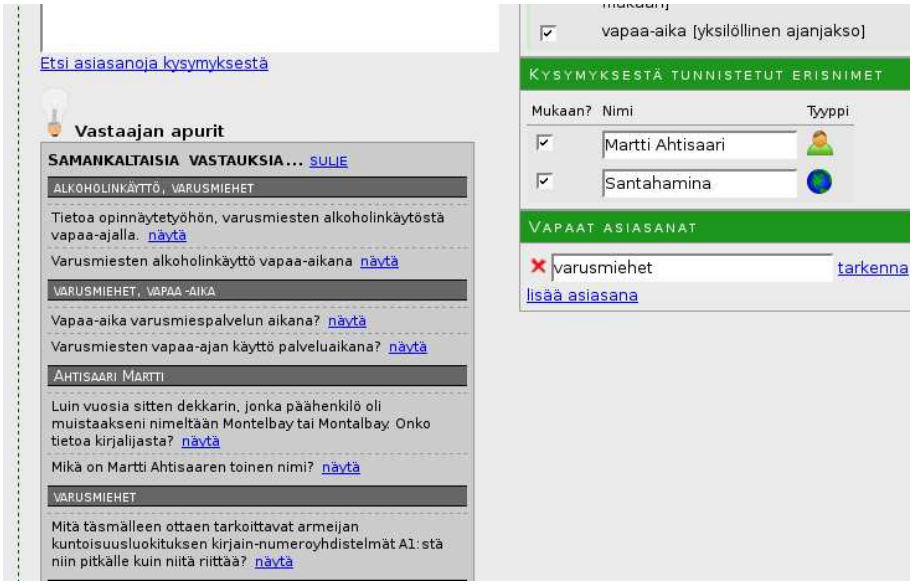
Kuva 7.6: Asiasanan tarkennus

7.4.3 Vastaajan apureiden käyttö

Asiasanojen valitsemisen jälkeen seuraa itse vastauksen kirjoittaminen. Tässä esimerkissä kirjastonhoitaja käyttää kaikkia vastaajan apureita vastauksen laatimisessa. Vastaajan apurit näkyvät kuvassa 7.4 kysymystekstin alla. Oletuksena kukin apuri on piilossa, ja *Näytä*-linkkiä klikkaamalla apurin saa esiin.

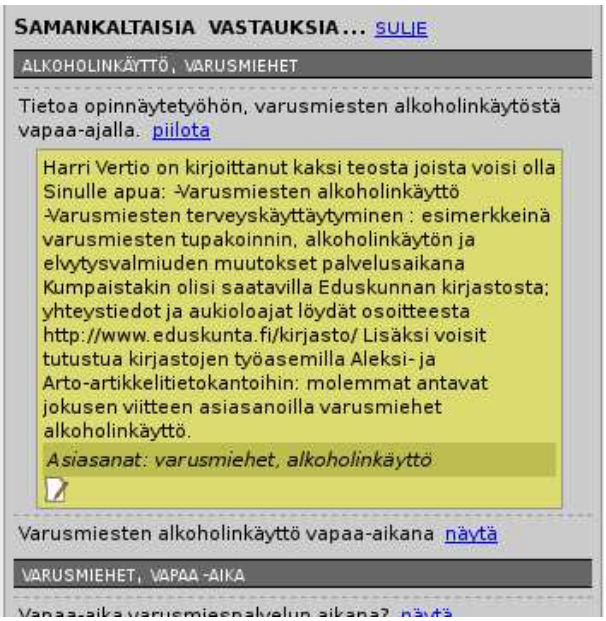
Samankaltaiset kysymykset

Samankaltaisten kysymysten hakijan tarkoituksena on näyttää kirjastonhoitajalle vanhoja kysymyksiä ja vastauksia, jotka mahdollisesti auttavat uuden vastauksen laatimisessa. Esimerkin hakutuloksia on havainnollistettu kuvassa 7.7. Tulokset on järjestetty sen mukaan, miten moni asiasana osuu niihin.



Kuva 7.7: Samankaltaisten kysymysten hakukomponentti

Kuvassa 7.8 kirjastonhoitaja on avannut yhden vanhan kysymyksen katsoakseen, onko se olennainen kysymyksen kannalta. Vastauksen alalaidassa olevalla paperia ja kynää esittävällä painikkeella kirjastonhoitaja voi liittää vanhan vastauksen uuden vastauksen pohjaksi.



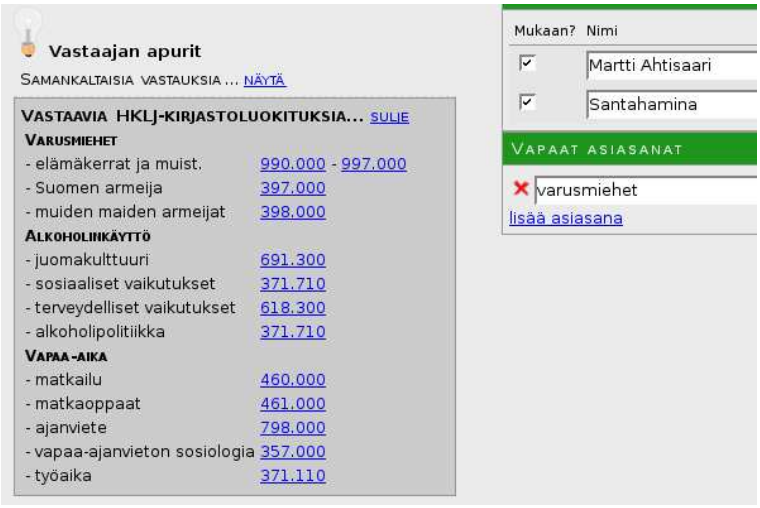
Kuva 7.8: Samankaltaisten kysymyksen katselu Oppaassa

HKLJ-luokitusten käyttö

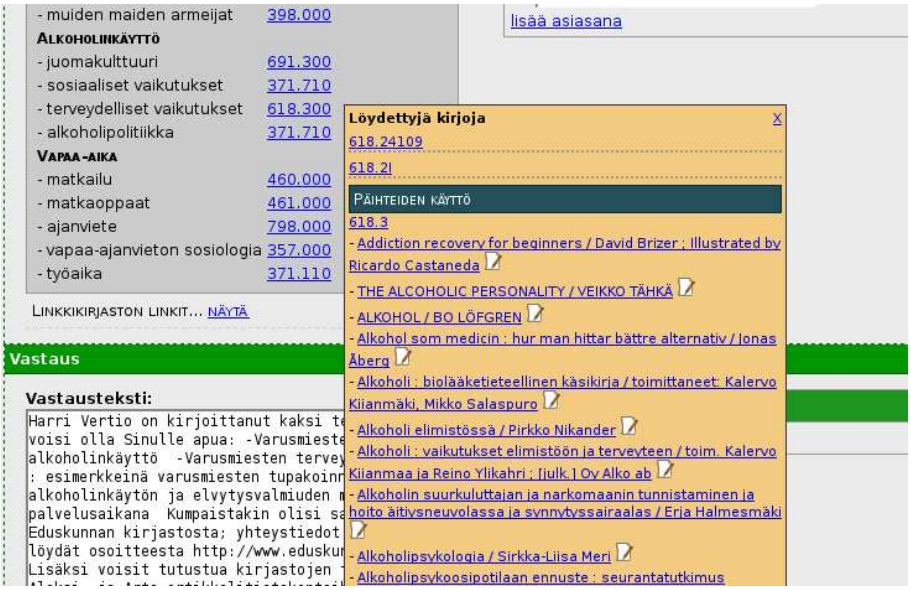
Kuvassa 7.9 on esitetty asiasanaehdotuksia vastaavia HKLJ-luokituksia sekä näkymiä. Klikkaamalla luokittelujärjestelmän luokan numeroa kirjastonhoitaja voi katsella luokkaan liittyviä kirjoja. Kirjahaku tehtiin Helmet-järjestelmästä[13]. Kirjahaun tuloksia ja kirjastoluokituksia voidaan käyttää kahdella tavalla: kirjastonhoitaja voi 1) itse etsiä niiden perusteella vastauksia kysymykseen tai 2) liittää vastaukseen linkkejä kirjoihin tai luokituksiin.

Kuvassa 7.10 on kuvattu tilanne, jossa kirjastonhoitaja on klikannut kirjastoluokkaa *618.300* ja Opas on listannut Helmet-järjestelmästä löytyviä luokkaan liittyviä kirjoja. Painamalla valkoista painiketta kirjastonhoitaja voi liittää kirjan nimen ja linkin Helmet-järjestelmään vastaustekstiin. Kirjan nimen klikkaaminen avaa selainikkunan Helmet-järjestelmään hakutulosten kohdalle, jos kirjastonhoitaja haluaa selata tarkemmin järjestelmän kirjoja.

[13]http://www.helmet.fi/



Kuva 7.9: HKLJ-kirjastoluokituksen katselu Oppaassa



Kuva 7.10: Kirjahaun tulosten katselu Oppaassa

Linkkikirjaston käyttö

Kuvassa 7.11 on esitetty asiasanaehdotuksia vastaavia luokittain ryhmiteltyjä linkkikirjaston linkkejä. Kirjastonhoitaja on avannut yhden luokan ja vienyt kursorin *Päihdelinkki*-luokan päälle. Klikkaamalla linkkiä kirjastonhoitaja voi

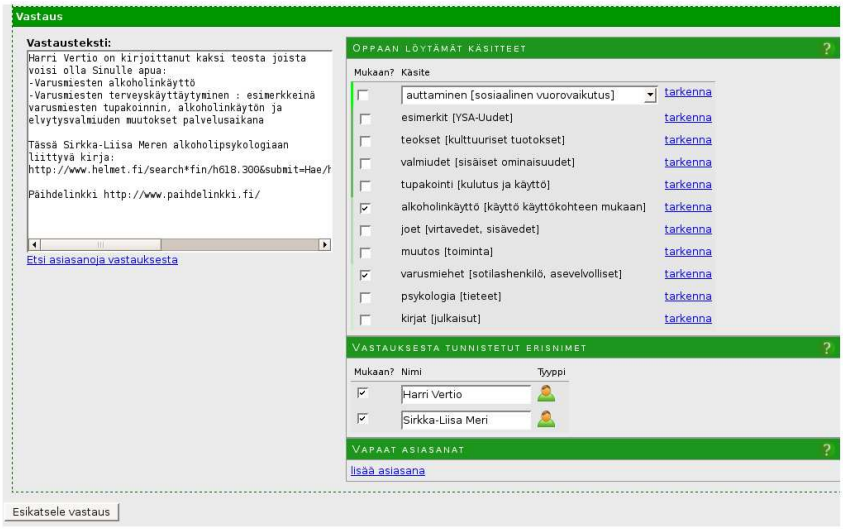
siirtyä linkin osoittamalle sivustolle. Valkoinen painike lisää linkin osoitteen vastaustekstiin.

Kuva 7.11: Linkkikirjaston linkkejä Oppaassa

7.4.4 Vastauksen muokkaus ja asiasanoitus

Kuvassa 7.12 nähdään, miten kirjastonhoitaja on vanhaa vastausta muokaten sekä ulkoisia resursseja (HKLJ, Linkkikirjasto) käyttäen kirjoittanut kysymykseen vastauksen. Myös vastauksesta voidaan etsiä asiasanoja painamalla *Etsi asiasanoja vastauksesta* -linkkiä. Tarkoituksena on, että jos vastauksessa esitellään esimerkiksi henkilöitä tai paikkoja, voidaan myös ne löytää automaattisesti ja liittää asiasanoitukseen.

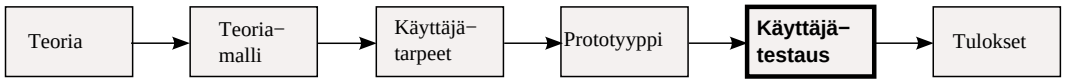
Esikatsele vastaus -painiketta painamalla päästään vastauksen esikatselutilaan ja edelleen vastauksen tallentamiseen kysymys-vastausarkistoon. Näitä toiminnallisuuksia ei toteutettu prototyyppiin.



Kuva 7.12: Vastauksen muokkaus ja asiasanoitus Oppaassa

Luku 8

Käyttäjätestit



Tässä kappaleessa esitellään edellisessä kappaleessa kuvatulla Oppaalla tehdyt käyttäjätestit.

8.1 Tavoitteet

Käytettävyystestit ovat olennainen osa ohjelmistokehitysprojektia, ja niitä kannattaa tehdä kaikissa projektin vaiheissa. Yksi käytettävyystestauksen alalajeista on käyttäjätestaus, jossa työstettävänä olevaa sovellusta testataan sovelluksen tulevien käyttäjien kanssa [DBT05]. Oppaan suhteen käyttäjätestit tehtiin siinä vaiheessa, kun ensimmäinen versio kirjastonhoitajan vastaamispuolesta tuli valmiiksi. Näiden käyttäjätestien tarkoituksena oli kartoittaa kirjastonhoitajien mielipiteitä Oppaan senhetkisestä tilasta, ja miten kirjastonhoitajat suhtautuvat uuteen, erilaiseen tapaan hoitaa kysymyksiin vastaaminen Kysy kirjastonhoitajalta -palvelussa.

8.2 Menetelmät

Käyttäjätesteissä käytettiin yhteisläpikäyntiä sekä jälkikäteen haastatteluja. Yhteisläpikäynnillä tarkoitetaan sitä, että käyttäjää ohjataan aktiivises-

ti käyttäjätestin aikana esimerkiksi kertoen prototyypin puutteista ja esitellen sen ominaisuuksia. Aktiiviseen osallistumiseen testin aikana päädyttiin siksi, koska Oppaassa oli niin paljon uusia ominaisuuksia olemassa olevaan sovellukseen verrattuna, ja koska Oppaassa oli jonkin verran toiminnallisia puutteita. Testihenkilöt laativat vastaustekstin kuitenkin itsenäisesti ilman aktiivista osallistumista. Jälkikäteen haastatteluja käytettiin testin jälkeen, jotta voitiin vielä rauhassa kysyä, minkälaisia ongelmia Oppaan käytössä ilmeni. Haastattelun pohjana olleet kysymykset ovat liitteessä 2.

8.3 Testatut käyttäjät

Benyon ym. [DBT05] suosittelee käyttäjätestejä varten kolmesta viiteen testikäyttäjää kustakin käyttäjäryhmästä. Työ keskittyi kirjastonhoitajan puoleen sovelluksessa, joten käyttäjätestejä tehtiin vain kirjastonhoitajien kanssa. Myös kirjastonhoitajien joukossa oli tunnistettavissa erilaisia käyttäjäryhmiä, mutta käyttäjätestien valossa heidät nähtiin kuitenkin yhteinäisenä ryhmänä.

Tutkimusta varten tehtiin neljä käyttäjätestiä yhteensä viiden käyttäjän kanssa. Yhdessä testitilanteessa oli kaksi kirjastonhoitajaa läsnä, mitä ei testeissä yleensä kannattaisi tehdä (esimerkiksi [DBT05]), mutta työn kannalta arveltiin, että oli parempi saada useamman ihmisen mielipiteitä Oppaasta.

Testihenkilöt olivat samasta joukosta, jonka kanssa tehtiin haastatteluja käyttäjätarpeiden määrittämiseksi:

- Yksi palvelun ylläpitäjä Kirjastot.fi-toimituksessa (nainen)
- Yksi palvelun valtakunnallinen ylläpitäjä (nainen)
- Kolme kirjastonhoitajaa (kaksi naista, yksi mies)

8.4 Testien kulku

Testien kulku meni siten, että ensin testihenkilölle kerrottiin, että kyse on Teknilliseen korkeakouluun tehtävästä diplomityöstä, ja että tarkoituksena on kartoittaa kirjastonhoitajien mielipiteitä mahdollisista uusista Kysy kirjastonhoitajalta -palveluun tulevista ominaisuuksista. Tämän jälkeen nämä Oppaan ominaisuudet esiteltiin. Testihenkilöä pyydettiin käyttämään Opasta kysymykseen vastaamiseen, aivan kuin tekisivät sen työssään. Testiä varten

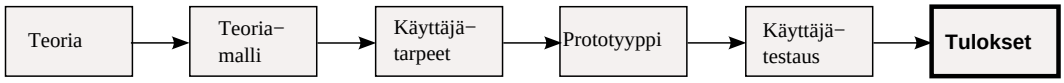
Oppaaseen oli syötetty seitsemän Kysy kirjastonhoitajalta -palveluun tullutta kysymystä:

- Voiko albaniankielisiä lastenkirjoja kaukolainata muista kirjastoista tai onko niitä ylipäätänsä miten saatavilla? Lapseni ovat kaksikielisiä (suomi ja albania).
- Haluaisin Thaimaasta Suomeen muuttaneelle vaimolleni jotain yksinkertaisia kieliharjoitteita. Vaikeutena on ollut löytää riittävä helppoja harjoitteita. Mitä suosittelisitte ja mitä käytetään?
- Mikä eläinlaji Suomessa on suurimmassa vaarassa kuolla sukupuuttoon???
- Etsin kirjasuosituksia ruotsinkieliselle 10-11 -vuotiaalle tytölle.
- Pari kirjavinkkiä esikouluikäiselle tarkoitetusta kirjasta, joissa aiheena on kaveruus ja ryhmässä toimiminen
- Miten leikitään nätisti yhdessä? Etsin iltatarinoiksi melkein 4-vuotiaalle kirjoja, joissa kerrotaan, miten kavereiden kanssa leikitään kauniisti ja kuinka kaverit saadaan pidettyä.
- Mikä kirja jossa olisi hauskaa ajanvietettä aikuisille?

Testihenkilöä pyydettiin valitsemaan näistä yksi vastattavaksi. Kysymykseen vastaamiseen testihenkilöillä meni noin kymmenestä kahteenkymmeneen minuuttia, riippuen miten paljon Oppaan ominaisuuksia käytiin läpi testin aikana. Testihenkilöt myös kertoivat oma-aloitteisesti sellaisia mielipiteitä, joita oli tarkoitus tiedustella vasta jälkikäteen haastattelussa. Kokonaisuudessaan kuhunkin testiin meni 45-60 minuuttia.

Luku 9

Tulokset



Tässä kappaleessa esitellään tutkimuksen tulokset. Tulosten esittely jakaantuu käyttäjätestien tuloksiin sekä työn aikana rakennetun prototyypin analyysiin käyttäjätestien tulosten valossa. Lopuksi vastataan johdannossa esitettyihin tutkimuskysymyksiin käyttäjätestien tulosten ja Oppaan laatimisessa kertyneiden kokemusten perusteella.

9.1 Käyttäjätestien tulokset

Yleensä ottaen testihenkilöt suhtautuivat myönteisesti prototyyppiin. Jokainen olisi valmis ottamaan prototyypinomaisen palvelun käyttöön, jos se olisi mahdollista. Seuraavassa on käsitelty käyttäjätestien tuloksia prototyypin ominaisuuksien mukaan jaoteltuna.

9.1.1 Asiasanojen ehdottaja

Asiasanojen ehdottajaa pidettiin hyväksyttävänä ominaisuutena ja asiasanoitustapaa selkeänä. Esiin tuli kuitenkin se ongelma, että asiasanaehdotuksia tulee liikaa, ja että ne voivat olla epäolennaisia, pitipä yksi testihenkilö asiasanaehdotuksia täysin epäolennaisina. Jaottelua käsitteiden, erisnimien ja vapaiden asiasanojen välillä pidettiin selvänä. Kukaan testihenkilöistä ei

käyttänyt testitehtävässä tarkenna/yleistä-toiminnallisuutta, ja siihen suhtauduttiin neutraalisti, ei torjuvasti eikä kehuenkaan. Selkeimmäksi puutteeksi asiasanaehdotuksissa nousi se, ettei prototyypissä voinut poistaa epäolennaisia asiasanoja.

Vapaiden asiasanojen täydentäjää (katso 7.4.2) pidettiin hyvänä ominaisuutena. Esiin tuli kuitenkin, että täydentäjän pitäisi toimia nopeasti, jotta kirjastonhoitajat eivät tuskastuisi sen käyttöön. Täydentäjän hyviksi puoliksi mainittiin kirjoitusvirheiden väheneminen ja se, ettei tarvitse erikseen avata asiasanasivustoa sanojen valintaa varten.

Sitä, että kysymyksen ja vastauksen asiasanat annetaan erikseen, ei pidetty kovinkaan hyvänä ominaisuutena, vaan pidettiin lähinnä ylimääräisenä vaivana. Sinällään oli hyväksyttävää, että myös vastaustekstistä etsitään mahdollisia asiasanoja. Yksi testihenkilö oli sitä mieltä, että erillinen asiasanoitus on hyväksyttävää.

Yksi testihenkilöistä arveli, että joissain tapauksissa asiasanoitustapa on jopa liian johdatteleva, jolloin vastauksia laaditaan asiasanojen perusteella. Hän oli myös sitä mieltä, ettei rutinoitunut vastaaja välttämättä tarvitse asiasanaehdotuksia, ja ne voivat jopa olla loukkaavia kirjastonhoitajan ammattitaitoa kohtaan.

9.1.2 Vastaajan apurit

Vastaajan apurit ja etenkin samankaltaisten kysymysten etsijä sai selkeästi positiivisimman vastaanoton Oppaassa, ja ajatusta vastaajan apureista pidettiin hyvänä. Myös sitä, että vastaajan apureista voi liittää suoraan vastaukseen tekstiä, pidettiin hyödyllisenä.

Etenkin HKLJ-kirjastoluokituksissa ja Linkkikirjastossa tuli esiin, että luokituksia ja linkkejä tuli liikaa. Kirjastoluokituksista löytyi kyllä vastaava luokka, mutta vasta pienen etsinnän jälkeen. Linkkikirjaston linkkejä vaivasi sama ongelma; usean löytyneen linkin seasta ei meinannut löytyä olennaisia linkkejä. Yksi vastaaja toivoi, että Linkkikirjaston hierarkiaa voisi lähteä selaamaan alaspäin apureissa, ja että prototyypissä olisi linkki Linkkikirjaston websivustoon.

Seuraavassa on listattu palveluja, joita ehdotettiin lisättäväksi vastaajan apureihin:

- Yleinen kirjastojen luokittelujärjestelmä YKL. (Oppaassa oli vain Hel-

singissä käytössä oleva HKLJ.)

- Makupalat[1]-linkkipalvelu, jota kirjastonhoitajien mainittiin käyttävän vastaustyössä. Osa vastaajista tosin piti Linkkikirjaston luokittelua parempana ja hyödyllisempänä.
- Kirjastojen käytettävissä olevat aineistotietokannat.
- Tiedonhaun portti[2], joka mainittiin hyödylliseksi erityisesti erilaisten lähteiden etsimiseen.
- Frank-monihaku[3], jonka avulla kirjahakuja voisi tehdä alueellisiin kirjastoihin.
- Linda[4]-, Helka[5]- ja Fennica[6]-tietokannat.

9.1.3 Tulosten luotettavuus

Tutkimuksen käyttäjätestit tehtiin suhteellisen pienelle joukolle, ja Benyon [DBT05] muistuttaakin, että siinä missä käyttäjätesteissä usein tulee paljon ideoita ja ongelmakohtia esiin, ei niistä kannata suinpäin tehdä yleistyksiä. On myös mahdollista, että testihenkilöt eivät vastanneet totuudenmukaisesti vaan myöntäillen, koska tiesivät, että testien tekijä oli sama henkilö, kuin joka oli ollut kehittämässä Opasta.

Kun kirjastonhoitajat vastaavat järjestelmään tuleviin kysymyksiin, he käyttävät usein käsikirjastoa vastaamisen apuna, ja saattavat kysyä kollegoiltaan apua. Osa käyttäjätesteistä jouduttiin tekemään muualla kuin kirjastossa, jossa kirjastonhoitajat yleensä vastaavat kysymyksiin. Tästä ja testitilanteesta johtuen testihenkilöt eivät luultavasti pystyneet toimimaan kysymykseen vastaamisessa aivan niin, kuin olisivat vastaamassa Kysy kirjastonhoitajalta -palvelua käyttäen. Oli myös havaittavissa, että testaustilanteesta johtuen testihenkilöt eivät vastanneet kysymykseen aivan samanlaisella huolellisuudella, kuin mitä ehkä oikeassa tilanteessa olisivat vastanneet.

[1] `http://www.makupalat.fi/`
[2] `http://tiedonhaunportti.kirjastot.fi/`
[3] `http://monihaku.kirjastot.fi/`
[4] `http://www.lib.helsinki.fi/kirjastoala/linnea/LINDA.htm`
[5] `http://www.helsinki.fi/helka/`
[6] `https://fennica.linneanet.fi/`

9.1.4 Johtopäätökset käyttäjätesteistä

Käyttäjätestien päätavoite oli selvittää, ollaanko Oppaan suhteen etenemässä oikeaan suuntaan ja millaisia mielipiteitä kirjastonhoitajilla on Oppaan vastaamispuolesta. Testien perusteella voidaan vetää johtopäätös, että puoliautomaattinen asiasanoitus ja vastaajan apurit olisivat hyödyllisiä ominaisuuksia Kysy kirjastonhoitajalta -palvelussa. Muutama seikka on kuitenkin hyvä muistaa. Ensinnäkin, prototyypissä tuli monta kertaa esiin, miten jokin ominaisuus oli hieman keskeneräinen ja toimi hieman hitaasti. Tällaisia keskeneräisyyksiä käyttäjät eivät siedä hyvin, vaan ne pitäisi saada minimoitua ja sovelluksen tulisi toimia sujuvasti. Toinen muistettava seikka on se, että Kysy kirjastonhoitajalta -palvelulla on pitkä historia takanaan ja sen käyttäjät ovat tottuneet tietynlaiseen vastaustapaan. Tästä syystä, kuten haastatteluissa ja testeissäkin tuli esiin, sovelluksen ominaisuuksien tulee olla hienovaraisia ja kunnioittaa kirjastonhoitajan ammattitaitoa. Esimerkiksi asiasanaehdotukset voivat olla aluksi piilossa ja kirjastonhoitaja niin halutessaan saa ne esiin.

Testihenkilöt hämmentyivät siitä, että vastaajan apureissa tuli niin paljon erilaisia resursseja, jotka ehkä liittyivät käsillä olevaan kysymykseen. Vaikka on houkutteleva ajatus, että kirjastonhoitajalle esitetään mielenkiintoisia tietolähteitä, joita voisi käyttää vastaamisessa apuna, vaikuttaa kuitenkin siltä, että on parempi näyttää vähän resursseja, jotka ovat kuitenkin oleellisia, kuin paljon resursseja, jotka *ehkä* ovat oleellisia kysymykseen nähden.

9.2 Oppaan suhde teoreettiseen malliin

Opas toteutettiin kappaleessa 5 esitetyn teoreettisen mallin perusteella. Oppaan käyttämä aineisto ja sen toiminnallisuudet soveltuivat hyvin tapauksiin perustuvan päättelyn askeleisiin, ja kirjastonhoitajan rooli nivoutui hyvin malliin. Sanasto-ontologiana käytetty Yleinen suomalainen ontologia toimi kätevänä “ontologisena liimana” Linkkikirjaston, Helsingin kirjaston luokittelujärjestelmän ja Oppaan kysymys-vastausaineiston välillä.

Oppaassa on kuitenkin jotain puutteita esitettyyn malliin nähden. Samankaltaisten kysymysten haku toteutettiin Oppaassa hyvin yksinkertaisesti ja se kohdistui pelkkiin vanhojen kysymys-vastausparien asiasanoihin. Lisäksi haussa ei hyödynnetty YSO:ssa määriteltyjä semanttisia suhteita, joilla hakutuloksia voitaisiin laajentaa ja parantaa. Asiasanaehdotuksissa ei myöskään hyödynnetty tietoa niin sanotuista implisiittisistä semanttisista klikeistä, eli siitä, minkälaisia asiasanoja on yleensä käytetty yhdessä Kysy kirjastonhoita-

jalta -palvelussa.

Ulkoisten aineistojen (Linkkikirjasto ja HKLJ) linkki YSO:on ja sitä kautta Oppaaseen tehtiin "heikosti", eli Linkkikirjaston ja HKLJ:n asiasanoilla ei ole eksplisiittistä suhdetta YSO:n käsitteisiin, vaan vertailu tehtiin käsitteiden nimien perusteella.

9.3 Jatkokehitysideoita

Opas oli ensimmäinen versio siitä, millainen vastaamispuoli Kysy kirjastonhoitajalta -palvelussä voisi olla. Sekä Opasta toteutettaessa että käyttäjätestien aikana ilmeni lukuisia kehitysideoita ja ongelmia. Seuraavassa on esitetty kehitysehdotuksia näiden ideoiden ja ongelmien perusteella. Ehdotukset on luokiteltu prototyypin toiminnallisuuksien mukaan.

9.3.1 Asiasanoitus

Tärkein kehityskohde asiasanoituksessa on se, miten Pokan ehdottamat asiasanat järjestetään ja karsitaan siten, että kysymys-vastausparin kannalta oikeasti olennaiset asiasanat tulisivat listalla ensimmäisinä. Ongelmallinen tilanne on erityisesti silloin, kun kysymys tai vastaus on pitkä, jolloin asiasanaehdotuksia tulee liikaa. Tällöin puoliautomaattisen asiasanoituksen edut hämärtyvät, kun asiasanojen keksimisen vaivasta tulee vain asiasanojen valinnan vaiva.

Asiasanat valittiin Oppaassa *Mukaan?*-valintalaatikon avulla. Intuitiivisempaa olisi, jos asiasanat olisivat oletuksena mukana ja huonot ehdotukset voisi poistaa samaan tapaan kuin vapaat asiasanat (katso 7.4.2).

Asiasanaehdotuksiin voitaisiin myös liittää enemmän tietoa asiasanaa vastaavan käsitteen ympäristöstä ontologiassa. Tällä hetkellä Oppaassa asiasanaehdotusten yhteydessä näytettiin käsitteen yläkäsite, joka testien mukaan useimmiten riitti erottelemaan käsite muista samannimisistä käsitteistä. Tämän lisäksi voitaisiin kuitenkin esimerkiksi vietäessä kursori käsitteen päälle näyttää enemmän käsitteen ympäristöä ontologiassa.

Automaattista semanttista täydentäjää voitaisiin kehittää edelleen. Vietäessä kursori täydentäjän ehdottamien käsitteiden päälle Oppaassa näytetiin "lisätietoikkunassa" käsitteen ympäristöä ontologiassa (katso 7.4.2), mutta tämän lisätietoikkunan kautta ei voitu valita käsitteitä. Tätä voitaisiin kehittää siten, että lisätietoikkunan kautta voitaisiin valita käsitteitä asiasanoik-

si ja mahdollisesti liikkua YSO:ssa edelleen muihin käsitteisiin.

Oppaan käyttöliittymästä puuttui tapa lisätä haluttaessa uusia paikkoja tai henkilöitä asiasanoitukseen. Oppaassa ei myöskään ollut tapaa erotella mitenkään paikkoja tai henkilönnimiä. Tähän on kaksi ratkaisutapaa, joita voidaan käyttää myös yhdessä: 1) henkilön ja paikan tyyppiä voisi vaihtaa. Tällöin esimerkiksi samannimiset mutta eri ammatissa toimivat ihmiset voitaisiin erotella eri tyypillä. Esimerkiksi Michael Jackson -asiasanan tyypiksi voitaisiin antaa YSO:n *viihdetaiteilijat* tai *asiantuntijat* riippuen siitä, onko kyse tunnetusta laulajasta vai oluiden asiantuntijasta. 2) Henkilöön ja paikkaan voisi liittää vapaamuotoisen erottelevan kommentin.

Oppaassa toteutettiin niin sanottua vapaata asiasanoitusta, jossa asiasanat valittiin yleisestä sanasto-ontologiasta (YSO) vapaasti. Jatkossa voitaisiin pohtia, olisiko löydettävissä jokin annotaatioskeema, jonka käyttäminen yhtenäistäisi asiasanoitusta ja mahdollistaisi paremmin näkymäpohjaisen haun toteuttamisen.

Oppaassa asiasanoitus tehdään erikseen kysymykselle ja vastaukselle. Hypoteesinä oli, että tapauksiin perustuvan päättelyn hakutoiminto toimisi paremmin, kun samankaltaisia kysymyksiä etsittäessä kysymys ja vastaus on asiasanoitettu erikseen. Tätä hypoteesiä ei kuitenkaan testattu, ja voi olla, että kysymyksen ja vastauksen asiasanoitus erikseen aiheuttaa turhaa monimutkaisuutta. Vaikka erottelu kysymys- ja vastausasiasanojen välillä säilytettäisiinkin, niin se pitäisi piilottaa käyttäjältä.

9.3.2 Vastaajan apurit

Vastaajan apureissa tärkeimmät konkreettiset kehityskohteet ovat seuraavat:

1. Samankaltaisten kysymysten hakemisessa haku kohdistetaan myös kysymystekstiin pelkkien asiasanojen sijasta.
2. Hyödynnetään asiasanojen YSO:ssa määriteltyjä semanttisia suhteita (*koirat*-haku osuu myös *terrierillä* asiasanoitettuihin kysymyksiin).
3. Linkkikirjastoon lisätään mahdollisuus selata linkkihierarkiaa.
4. Linkkikirjaston vastausjoukon luokittelua pitäisi miettiä uudelleen, testien mukaan jako Linkkikirjaston luokkien mukaan ei vaikuttanut hyvältä.

5. Vastaajan apureiden sisältö päivittyy dynaamisesti koko ajan taustalla käyttäjän tehdessä asiasanavalintoja eikä vasta silloin, kun käyttäjä avaa jonkin apurin.

Vastaajan apureissa pitäisi miettiä myös sitä, että on olemassa oikeastaan kahdentyyppisiä ulkoisia aineistoja: 1) aineistoja, joita voi käyttää "inspiraation lähteenä" ja 2) resursseja, joita voi sellaisenaan liittää vastaukseen. Esimerkiksi samankaltaiset vastaukset voivat auttaa vastaamistyössä, vaikka vastaus ei osuisikaan aivan yhteen esitetyn kysymyksen kanssa. Sen sijaan Linkkikirjaston linkeissä testien mukaan vaikutti siltä, että tulosten pitää olla tarkempia, jotta niistä on hyötyä vastaamistyössä. "Sinne päin" olevista linkeistä oli Oppaassa vain haittaa. Tämä vastaajan apureiden kaksijakoisuus pitäisi ottaa huomioon Oppaan jatkokehityksessä esimerkiksi siten, että Linkkikirjastossa asiasanahaussa käytetään JA-relaatiota kun taas samankaltaisten kysymysten haussa TAI-relaatiota.

9.3.3 Vastauksen kirjoittaminen

Oppaassa kysymys- ja vastaustekstin kirjoittaminen tapahtui yksinkertaisilla HTML-tekstialuekomponenteilla. Tässä voitaisiin hyödyntää niin sanottuja rikkaita tekstieditoreita, jotka mahdollistavat tekstinkäsittelyohjelmista tuttuja toiminnallisuuksia, kuten tekstin lihavoinnin, kursivoinnin ja kuvien liittämisen. Yksi tällainen rikas tekstieditori on *Dojo Rich Text Editor*[7].

9.3.4 Yleisiä havaintoja Oppasta

Aineistojen ontologisointi

Aineistojen ontologisoinnissa ongelmallista oli eri asiasanojen suhteuttaminen YSO-käsitteisiin. Etenkin kysymys-vastausaineiston asiasanoissa oli paljon kirjoitusvirheitä eikä voitu tietää, tarkoitetaanko jollain asiasanalla jotain tiettyä YSO:n käsitettä. Tehokkaan menetelmän kehittäminen olemassa olevien aineistojen ontologisointiin ei ollut työn tavoitteena, joten liittäminen tehtiin automaattisesti merkkijonovertailua käyttämällä.

Oppaassa jätettiin myös hyödyntämättä aineistossa olevat nimet, joita etenkin kysymys-vastausaineistossa oli paljon. Pokan nimentunnistajaa oltaisiin voitu

[7] `http://dojotoolkit.org/docs/rich_text.html`

käyttää näiden tunnistamiseen. Poka tuottaa kuitenkin paljon myös virheellisiä nimitunnistuksia, ja aineiston laajuuden vuoksi niitä ei haluttu käydä manuaalisesti läpi.

Opas hyötyisikin siitä, että asiasanojen suhde YSO:n käsitteisiin käytäisiin manuaalisesti läpi, jolloin kirjoitus- ja muut virheet voitaisiin korjata ja puuttuvat suhteet lisätä. Sama koskee kysymys-vastausaineistossa olevia nimiä.

Tietomalli

Oppaan käyttämä kysymys-vastausontologia ja sen suhteutus ulkopuolisiin aineistoihin, etenkin YSO:on on toteutettu kysymyksiin vastaamisen näkökulmasta, ja toteutettu hakutoiminnallisuus on hyvin yksinkertainen. Voikin olla, että esimerkiksi loppukäyttäjän versiota toteutettaessa tai hakutoiminnallisuutta kehitettäessä tätä tietomallia joudutaan muuttamaan.

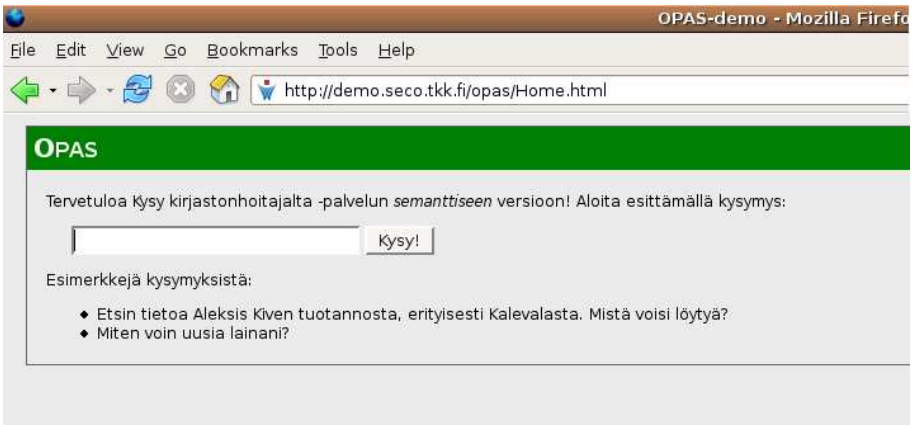
Asiasanojen tarkentaminen (katso 7.3.1) ei tällä hetkellä tee erottelua, onko tarkennus jonkin käsitteen alakäsite vai ilmentymä. Oppaassa ei katsottu, että tarkennuksia tulisi niin paljon, että erottelu olisi tarpeen, mutta jatkokehityksessä erottelun tekeminen voi olla tarpeellista.

9.3.5 Ideoita loppukäyttäjän versiosta

Loppukäyttäjät käyttävät Kysy kirjastonhoitajalta -palvelua luultavasti kahdesta syystä: 1) heillä on jokin kysymys, jolloin he ehkä selaavat arkistoa löytääkseen aiheeseen liittyviä kysymyksiä tai 2) he selailevat palvelun arkistoa mielenkiinnosta. Nämä kaksi kohtaa on pidettävä mielessä loppukäyttäjän versiota kehitettäessä.

Samankaltaisten kysymysten etsijää voisi hyödyntää loppukäyttäjän versiossa lähes sellaisenaan ainakin kahdella tavalla: 1) kun käyttäjä on lähettämässä kysymystä palveluun, samankaltaisten kysymysten etsijällä haetaan vastaavia kysymyksiä ja ehdotetaan käyttäjälle “Olisiko näissä kysymyksissä jo vastaus kysymykseesi?” 2) arkiston hakutoimintoa muutetaan siten, että käyttäjät voivat esittää ongelmansa luonnollisella kielellä, ja samankaltaisten kysymysten etsijää käytetään näyttämään käyttäjän pulmaan liittyviä kysymyksiä. Kuvassa 9.1 on hahmoteltu tällaisen käyttöliittymän alkunäkymää.

Loppukäyttäjän versiossa myös vastaajan apureita voitaisiin käyttää näyttämään kysymykseen liittyviä mielenkiintoisia resursseja, kun käyttäjä on valinnut jonkin kysymys-vastausparin nähtäväkseen. Vastaajan apurei-



Kuva 9.1: Käyttöliittymähahmotelma loppukäyttäjän versiosta Oppaassa

den sisältöä jouduttaisiin luultavasti muuntamaan jonkin verran, koska todennäköisesti loppukäyttäjällä on erilaiset tiedontarpeet kuin kirjastonhoitajalla.

Loppukäyttäjän versiossa voitaisiin myös miettiä, miten Oppaassa voitaisiin soveltaa sitä, että käyttäjän haku pitäisi nähdä prosessina eikä vain yksittäisenä toimenpiteenä. Loppukäyttäjä pitäisikin saada mukaan CBR-syklin hakuvaiheeseen. Esimerkkinä vaikkapa tilanne, jossa käyttäjä on kirjoittamassa hakusanaa *Michael Jackson*, ja kun järjestelmä tunnistaa nimen, ja ennen kuin käyttäjä painaa *Hae*-nappia, järjestelmä pyytää erottelemaan artisti-Jacksonin ja olutasiantuntija-Jacksonin välillä.

9.3.6 Oppaan hyödyntäminen ilman semanttista tiedoneristäjää

Olennainen osa Opasta on semanttinen tiedoneristyskomponentti *Poka*. Voi kuitenkin olla, että syystä tai toisesta työtä hyödyntävällä taholla ei ole käytettävissä Pokaa tai muuta vastaavaa semanttista tiedoneristäjää. Tässä tapauksessa Oppaasta jää puuttumaan vain asiasanaehdotukset, mutta esimerkiksi vapaiden asiasanojen lisääminen ja automaattinen semanttinen täydennys toimii.

9.4 Vastaukset tutkimuskysymyksiin

Projektin alussa työlle asetettiin kaksi tutkimuskysymystä. Ensimmäinen tutkimuskysymys kuului:

> *M*iten semanttisen webin tekniikoita voidaan hyödyntää helpottamaan vastaamistyötä kysymys-vastauspalveluissa?

Suurin osa työstä keskittyi tähän kysymykseen, koska työn alussa Kirjastot.fi-toimituksen kanssa käytyjen keskustelujen perusteella juuri vastaamistyön helpottaminen koettiin tärkeäksi kehityskohteeksi. Vastaus tähän kysymykseen jakaantuu hyötyihin itse vastaamistyössä ja hyötyihin metatiedon liittämisessä, asiasanoituksessa:

- Semanttista tiedoneristämistä voidaan käyttää asiasanaehdotusten luomiseen, mikä auttaa kysymys-vastausaineiston asiasanoituksessa.
- Ontologioissa määriteltyjä semanttisia suhteita voidaan käyttää asiasanaehdotusten oletetun relevanssin selvittämiseen, mikä osaltaan auttaa kysymys-vastausaineiston asiasanoituksessa.
- Ontologioita voidaan käyttää "semanttisena liimana" erilaisten aineistojen välillä, jolloin eri aineistoja voidaan helpommin yhdistää ja käyttää vastauksen kirjoittamisessa.
- Ontologiapohjaista asiasanoitusta voidaan käyttää välttämään kirjoitusvirheitä ja yhtenäistämään aineistoa.
- Automaattista semanttista täydentämistä voidaan käyttää helpottamaan käsitteiden etsimiseen ontologioista.

Toinen tutkimuskysymys kuului:

> *M*illaisia uusia mahdollisuuksia semanttisen webin tekniikat avaavat kysymys-vastauspalveluiden toteuttamisessa?

Vastaukset tähän tutkimuskysymykseen perustuvat Oppaan toteuttamisessa saatuihin kokemuksiin:

- Ontologioita voidaan käyttää tietomallina erilaisissa kysymys-vastauspalveluissa.
- Vastaamistyön ohella myös yleisellä tasolla ontologiat toimivat semanttisena liimana erilaisten aineistojen välillä, jolloin eri aineistoja voidaan helposti yhdistää kysymys-vastauspalveluihin.
- Ajax-tekniikkaa kannattaa käyttää kysymys-vastauspalveluiden käyttöliittymätoiminnallisuuksien tekemisessä.

Näistä Ajax ei ole semanttisen webin tekniikka sinänsä, mutta sitä käytetään kuitenkin usein semanttisen webin sovelluksissa.

Luku 10

Johtopäätökset

Tässä raportissa käsiteltiin sitä, miten semanttisen webin tekniikoita voidaan hyödyntää kysymys-vastauspalveluissa. Tämän tutkimiseksi työssä laadittiin malli tapauksiin perustuvaa päättelyä (engl. *Case-based Reasoning*) ja semanttista asiasanoitusta yhdistävästä kysymys-vastauspalvelusta. Tämän mallin perusteella rakennettiin *Opas*, semanttinen ontologioihin pohjautuva Kysy kirjastonhoitajalta -sovelluksen prototyyppi.

Työssä kävi ilmi, että tapauksiin perustuvan päättelyn teoriaa voidaan hyvin soveltaa kysymys-vastauspalveluiden toteuttamisessa. Semanttista tiedoneristystä voidaan käyttää kysymys-vastausparien asiasanoituksessa laatimalla syötetekstin perusteella asiasanaehdotuksia, jotka vastausten laatija valinnoillaan vahvistaa. Ontologiat osoittautuivat toimiviksi tietorakenteiksi sekä Oppaan tietomallina että erilaisten aineistojen yhdistämisessä. Voidaan päätellä, että semanttisen webin tekniikat soveltuvat hyvin kysymys-vastauspalveluiden toteuttamiseen.

Työstä avautui paljon jatkotutkimuksen lähtökohtia semanttisten kysymys-vastauspalveluiden parissa. Työssä ei tutkittu ollenkaan loppukäyttäjien puolta kysymys-vastauspalveluissa. Tässä mielenkiintoinen tutkimisen aihe on esimerkiksi se, miten vastauksia kannattaa kategorisoida ja minkälaisia näkymiä aineistoon voidaan luoda moninäkymähaun soveltamiseksi. Tutkimisen arvoista on myös se, miten monilähteistä semanttista suosittelua [Vil06] voitaisiin käyttää arkiston selailussa näyttämään mielenkiintoisia muita kysymyksiä ja resursseja.

Ontologioiden käyttäminen asiasanoituksessa yhdenmukaistaa asiasanoitusta, mutta ontologiat eivät silti suoranaisesti ohjeista minkäänlaiseen asiasanoitukseen. Kysymys-vastauspalveluissa voitaisiinkin jollain tapaa ohjata “noviiseja”

tekemään asiasanoitusta etevien asiasanoittajien tavoin. Etevämpiä asiasanoittajia voitaisiin tunnistaa esimerkiksi siten, miten usein heidän asiasanoittamat kysymykset osuvat arkistohaussa hakutuloksiin, ja miten usein näitä kysymyksiä klikataan. Etevyys voidaan myös määrittää eksplisiittisillä rooleilla, jos tiedetään tiettyjen asiasanoittajien olevan etevämpiä kuin toiset. Tässä voisi tutkia niin sanotun *yhteissuodatuksen* (engl. *collaborative filtering*) [HKR00] teoriaa, ja miten sitä voidaan yhdistää kysymys-vastauspalveluihin ja asiasanoitukseen.

Kysymyksiin vastaajien yhteistyötä ja asiasanoituksessa laatimia käsitteiden tarkennuksia voitaisiin käyttää myös asiasanoituksessa käytettävän sanaston laajentamiseen. Luotuja alakäsitteitä ei kannata lisätä suoraan mihinkään julkiseen sanastoon, mutta ne voivat toimia ehdotuksina uusista sanastokäsitteistä.

Kysy kirjastonhoitajalta -palvelussa kysymykset ohjataan kirjastonhoitajille kysyjän asuinpaikkakunnan mukaan. Tietoa kysyjän asuinpaikasta voitaisiin hyödyntää palvelussa myös muilla tavoin. Esimerkiksi vastaajan apureita voitaisiin kustomoida sen mukaan, missä kysymyksen esittäjä asuu. Kirjahaku voitaisiin kohdistaa asuinpaikkakunnan kirjastojärjestelmiin ja vastaajan apureissa painotettaisiin asuinpaikkakuntaan liittyviä tietolähteitä. Näin vastaajan apurit tarjoaisivat mahdollisesti kiinnostavampaa tietoa kysyjän kannalta.

Asuinpaikkakunnan lisäksi kirjastonhoitajat valitsevat kysymyksiä myös mielenkiinnon ja osaamisen mukaan. Erityisesti palvelussa mukana olevat 14 erikoiskirjastoa vastaavat erikoisaloihinsa liittyviin kysymyksiin. Kirjastonhoitajille ja vastaajakirjastoille voitaisiinkin laatia osaamista ja mielenkiintoa kuvaavat ontologiset profiilit. Kysymysteksistä louhittujen käsitteiden perusteella kysymyksiä voitaisiin ohjata näille profiileille ja yrittää päätellä, kuuluuko kysymys jollekin tietylle erikoiskirjastolle.

Oppaassa käytettiin niin sanottua vapaata asiasanoitusta, jossa asiasanat valitaan yleisestä sanasto-ontologiasta. Vaihtoehtona tähän voitaisiin tutkia SeCo-tutkimusryhmässä kehitettävän niin sanotun *kehyspohjaisen annotoinnin* käyttöä vapaan asiasanoituksen sijasta. Siinä syötetekstistä etsitään pelkkien käsitteiden sijaan *kehyksiä*, jotka koostuvat tekijästä, tekemisestä ja tekemisen kohteesta. Näin loppukäyttäjä voisi esimerkiksi hakea kysymyksiä, joissa jokin hakusana esiintyy juuri tietyssä roolissa (esimerkiksi *Picasso maalaa*) sen sijaan, että käyttäisi yksittäisiä hakutermejä (*Picasso*, *maalaus*).

Lähdeluettelo

[AKA91] D. W. Aha, D. Kibler, and M. K. Albert, *Instance-based learning algorithms*, Mach. Learn. **6** (1991), no. 1, 37–66.

[AP94] A. Aamodt and E. Plaza, *Case-based reasoning: foundational issues, methodological variations, and system approaches*, AI Commun. **7** (1994), no. 1, 39–59.

[Arm00] W. Arms, *Digital libraries, glossary*, 2000, `http://www.cs.cornell.edu/wya/DigLib/MS1999/glossary.html`.

[Av04] G. Antoniou and F. vanHarmelen, *A semantic web primer*, MIT Press, Cambridge, MA, USA, 2004.

[BHK+97] R. D. Burke, K. J. Hammond, V. A. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg, *Question answering from frequently asked question files: Experiences with the faq finder system*, Tech. report, University of Chicago, Chicago, IL, USA, 1997.

[Bic04] I. Bichindaritz, *Mémoire: Case based reasoning meets the semantic web in biology and medicine*, Lecture Notes in Computer Science **3155** (2004), 47–61.

[Bil04] M. W. Bilotti, *Query expansion techniques for question answering*, Master's thesis, Massachusetts Institute of Technology (MIT), May 2004.

[BLHL01] T. Berners-Lee, J. Hendler, and O. Lassila, *The semantic web*, Scientific American **284** (2001), no. 5, 28–37.

[BYRN99] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, ch. Introduction, pp. 1–3, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[CDPW02] F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks, *User-system cooperation in document annotation based on information extraction*, EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (London, UK), Springer-Verlag, 2002, pp. 122–137.

[CRCC96] K. H. Chang, P. Raman, W. H. Carlisle, and J. H. Cross, *A self-improving helpdesk service system using case-based reasoning techniques*, Comput. Ind. **30** (1996), no. 2, 113–125.

[CW03] H. Chen and Z. Wu, *Case markup language for case-based reasoning in semantic web*, 2003, `http://www710.univ-lyon1.fr/~bfuchs/ws-iccbr03/papers/06_Wu_Chen.pdf`, viitattu 5.10.2006.

[DBT05] P. Turner D. Benyon and S. Turner, *Designing interactive systems, people, activities, contexts, technologies*, p. 277, Pearson Education Limited, 2005.

[DEG+03] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien, *Semtag and seeker: bootstrapping the semantic web via automated semantic annotation*, WWW '03: Proceedings of the 12th international conference on World Wide Web (New York, NY, USA), ACM Press, 2003, pp. 178–186.

[DM02] S. Delisle and B. Moulin, *User interfaces and help systems: from helplessness to intelligent assistance*, Artif. Intell. Rev. **18** (2002), no. 2, 117–157.

[EMSS00] M. Erdmann, A. Maedche, H. Schnurr, and S. Staab, *From manual to semi-automatic semantic annotation: About ontology-based text annotation tools*, 2000.

[Fal03] G. Falkman, *Issues in structured knowledge representation a definitional approach with application to case-based reasoning and medical informatics*, Ph.D. thesis, Chalmers University of Technology, Göteborg University, 2003.

[FHLL00] S. Foo, S. C. Hui, P. C. Leong, and S. Liu, *An integrated help support for customer services over the world wide web: a case study*, Comput. Ind. **41** (2000), no. 2, 129–145.

[FLGD87] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, *The vocabulary problem in human-system communication*, Commun. ACM **30** (1987), no. 11, 964–971.

[FS03] G. Finnie and Z. Sun, *R5 model for case-based reasoning*, Knowledge-Based Systems **16** (2003), no. 1, 59–65.

[FSKH06] D. Feng, E. Shaw, J. Kim, and E. Hovy, *An intelligent discussion-bot for answering student queries in threaded discussions*, IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces (New York, NY, USA), ACM Press, 2006, pp. 171–177.

[GR96] A. R. Golding and P. S. Rosenbloom, *Improving accuracy by combining rule-based and case-based reasoning*, Artif. Intell. **87** (1996), no. 1-2, 215–254.

[GRB99] M. H. M. H. Goker and T.T. Roth-Berghofer, *The development and utilization of the case-based help-desk support system homer*, Engineering Applications of Artificial Intelligence **12** (1999), no. 6, 665–680.

[HBLH94] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, *Ohsumed: an interactive retrieval evaluation and new large test collection for research*, SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), Springer-Verlag New York, Inc., 1994, pp. 192–201.

[Hef] J. Heflin, *Owl web ontology language use cases and requirements*, `http://www.w3.org/TR/webont-req/#onto-def`, viitattu 4.10.2006.

[HHL06] M. Holi, E. Hyvönen, and P. Lindgren, *Integrating tf-idf weighting with fuzzy view-based search*, Proceedings of the ECAI Workshop on Text-Based Information Retrieval (TIR-06), Aug 2006.

[HKR00] J. L. Herlocker, J. A. Konstan, and J. Riedl, *Explaining collaborative filtering recommendations*, CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work (New York, NY, USA), ACM Press, 2000, pp. 241–250.

[HLB99] A. Hanjalic, R. Lagendijk, and J. Biemond, *Semi-automatic news analysis, indexing, and classification system based on topics preselection*, 1999.

[HM00] U. Hahn and I. Mani, *The challenges of automatic summarization*, Computer **33** (2000), no. 11, 29–36.

[HM06] E. Hyvönen and E. Mäkelä, *Semantic autocompletion*, Proceedings of the 1st Asian Semantic Web Conference (ASWC-2006), Beijing, Sep 4-9, 2006, forth-coming.

[HSV04] E. Hyvönen, S. Saarela, and K. Viljanen, *Application of ontology techniques to view-based semantic search and browsing*, The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004), 2004.

[HVK+05] E. Hyvönen, A. Valo, V. Komulainen, K. Seppälä, T. Kauppinen, T. Ruotsalo, M. Salminen, and A. Ylisalmi, *Finnish national ontologies for the semantic web - towards a content and service infrastructure*, Proceedings of International Conference on Dublin Core and Metadata Applications (DC 2005), Nov 2005.

[KEW01] C. C. T. Kwok, O. Etzioni, and D. S. Weld, *Scaling question answering to the web*, WWW '01: Proceedings of the 10th international conference on World Wide Web (New York, NY, USA), ACM Press, 2001, pp. 150–161.

[KK01] J. Kahan and M. Koivunen, *Annotea: an open rdf infrastructure for shared web annotations*, WWW '01: Proceedings of the 10th international conference on World Wide Web (New York, NY, USA), ACM Press, 2001, pp. 623–632.

[LGB06] G. Schreiber L. Gazendam, V. Malaisé and H. Brugman, *Deriving semantic annotations of an audiovisual program from contextual texts*, Semantic Web Annotation of Multimedia (SWAMM'06) workshop, 2006, `http://www.cs.vu.nl/~guus/papers/Gazendam06a.pdf`.

[LMU06] V. Lopez, E. Motta, and V. Uren, *Poweraqua: Fishing the semantic web*, Proceedings of ESWC 2006, 2006.

[Mil95] G. A. Miller, *Wordnet: a lexical database for english*, Commun. ACM **38** (1995), no. 11, 39–41.

[MPS98] D. L. McGuinness and P. F. Patel-Schneider, *Usability issues in knowledge representation systems*, AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence (Men-

lo Park, CA, USA), American Association for Artificial Intelligence, 1998, http://scholar.google.fi/url?sa=U&q=http://www.cs.umbc.edu/771/papers/D%Lusability.pdf, pp. 608–614.

[Mud98] M. R. Muddamalle, *Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics*, J. Am. Soc. Inf. Sci. **49** (1998), no. 10, 881–887.

[MvH] D. L. McGuinness and F. van Harmelen, *Owl web ontology language overview*, http://www.w3.org/TR/owl-features/.

[MZ05] Y. Marom and I. Zukerman, *Analysis and synthesis of help-desk responses*, Lecture Notes in Computer Science, vol. 3683, pp. 890–897, Springer Berlin / Heidelberg, 2005.

[RDS93] H. Shrobe R. Davis and P. Szolovits, *What is a knowledge representation?*, AI Magazine **14** (1993), no. 1, 17–33.

[RFQ+05] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal, *Probabilistic question answering on the web*, Journal of the American Society for Information Science and Technology **56** (2005), no. 6, 571–583.

[RH05] L. Reeve and H. Han, *Survey of semantic annotation platforms*, SAC '05: Proceedings of the 2005 ACM symposium on Applied computing (New York, NY, USA), ACM Press, 2005, pp. 1634–1638.

[SDWW01] G. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, *Ontology-based photo annotation*, IEEE Intelligent Systems **16** (2001), no. 3, 66–74.

[SG01] R. Schmidt and L. Gierl, *Cased-based reasoning for medical knowledge-based systems*, International Journal of Medical Informatics **64** (2001), no. 2, 355–367.

[Sim92] E. Simoudis, *Using case-based retrieval for customer technical support*, IEEE Expert: Intelligent Systems and Their Applications **7** (1992), no. 5, 7–12.

[SM86] G. Salton and M. J. McGill, *Introduction to modern information retrieval*, McGraw-Hill, Inc., New York, NY, USA, 1986.

[Sri96] P. Srinivasan, *Optimal document-indexing vocabulary for medline*, Inf. Process. Manage. **32** (1996), no. 5, 503–514.

[VAH06] A. Vehviläinen, O. Alm, and E. Hyvönen, *Combining case-based reasoning and semantic indexing in a question-answer service*, June 20 2006, Poster paper, 1st Asian Semantic Web Conference (ASWC2006).

[vAMS+04] M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga, *A method for converting thesauri to rdf/owl*, Third International Semantic Web Conference ISWC 2004, vol. 3298, 2004.

[VHA06] A. Vehviläinen, E. Hyvönen, and O. Alm, *A semi-automatic semantic annotation and authoring tool for a library help desk service*, Proceedings of the first Semantic Authoring and Annotation Workshop, November 2006, To be published.

[Vil06] K. Viljanen, *Monilähteinen suosittelu semanttisessa webissä*, Master's thesis, University of Helsinki, March 6 2006.

[Voo01] E. M. Voorhees, *Overview of the TREC 2001 question answering track*, Text REtrieval Conference, 2001.

[Wat98] I. D. Watson, *Applying case-based reasoning: techniques for enterprise systems*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[XC96] J. Xu and W. B. Croft, *Query expansion using local and global document analysis*, SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), ACM Press, 1996, pp. 4–11.

[Xri00] F. Xristine, *Usability engineering*, Palgrave, 2000.

LIITE 1. Käyttäjähaastattelujen kysymykset

Seuraavia kysymyksiä käytettiin käyttäjähaastattelujen pohjana.

Yleistä palvelusta

1. Kuvaile vapaasti, miten käytät Kysy kirjastonhoitajalta -palvelua työpäiväsi aikana.
2. Miten kauan olet käyttänyt palvelua?
3. Miten usein käytät palvelua?
4. Mainitse kolme hyvää puolta palvelusta.
5. Mainitse kolme huonoa puolta palvelusta.
6. Jos saisit vapaasti päättää kolme uutta ominaisuutta palveluun, mitkä ne olisivat?

Kysymyksiin vastaaminen

1. Kuinka teet päätöksen kysymykseen vastaamisesta?
2. Miten haet tietoa (Google, sisäiset tiedonlähteet, kollegat, Kysy kirjastonhoitajalta -arkisto jne.)
3. Miten kauan kysymykseen vastaamiseen menee aikaa, ja teetkö sen "kertarykäyksellä"vai pienemmissä palasissa?
4. Mikä on kysymykseen vastaamisessa helpointa? Entä mikä vaikeinta?

5. Mikä on näkemyksesi siihen, että palvelussa olevat kysymykset toistuvat ajan mittaan?

Vastausten asiasanoittaminen

1. Jokaiseen vastaukseen liitetään siihen liittyviä asiasanoja. Mitä mieltä olet tästä?
2. Mikä asiasanoittamisessa on helppoa? Entä vaikeata?
3. Onko olemassa tietynlaisia kysymyksiä, jonka vastauksiin on helppoa liittää asiasanoja? Entä vaikeaa?

Poikkeukselliset tilanteet

1. Kuvaile tilannetta, jossa vastaaminen on epäonnistunut (jos näin on käynyt). Mistä tilanne johtui? Mitä teit tilanteelle?
2. Onko käynyt niin, että katsoit kysymyksen kuuluvan sinulle, mutta et jostain syystä (ajanpuute? tiedonpuute?) pystynyt vastaamaan siihen?

LIITE 2. Käyttäjätestien jälkihaastattelun kysymykset

Seuraavia kysymyksiä esitettiin käytettävyystestauksen yhteydessä, kun testihenkilö oli suoriutunut tehtävistään.

Asiasanoittamisesta

1. Olivatko järjestelmän automaattisesti ehdottamat asiasanat mielestäsi kysymykseen tai vastaukseen liittyen olennaisia?
2. Oliko erottelu käsitteiden, erisnimien ja vapaiden asiasanojen välillä selkeä?
3. Jos lisäsit vapaita asiasanoja, mitä mieltä olet automaattisesta asiasanojen ehdottajasta?
4. Mikä on hyvää ja mikä huonoa sovelluksen asiasanoitustavassa?
5. Miten koet sen, että asiasanoitus tehdään erikseen kysymykselle ja vastaukselle?
6. Tuntuiko tarkenna/yleistä-toiminnallisuus intuitiiviselta?
7. Yleensä ottaen, tuntuiko asiasanoittaminen selkeältä sovelluksessa?

Vastaajien apurit

1. Jos käytit vastaamisessa vastaajan apureita, tuntuiko, että niistä oli vastaamistyössä apua?
2. Oliko samankaltaisissa vastauksissa mielestäsi oleellisia vastauksia uuteen kysymykseen liittyen? Oliko niitä helppo käyttää hyväksi?

3. Oliko HKLJ-luokituksissa mielestäsi oleellisia luokkia uuteen kysymykseen liittyen? Oliko niitä helppo käyttää hyväksi?

4. Oliko linkkikirjastossa mielestäsi oleellisia linkkejä uuteen kysymykseen liittyen? Oliko niitä helppo käyttää hyväksi?

5. Miten kehittäisit vastaajan apureita, jos sinulle annettaisiin siihen mahdollisuus?

6. Koetko tarpeelliseksi, että vastaajan apureista käsin voi lisätä vastaustekstiin automaattisesti tekstiä?

Yleisiä kysymyksiä

1. Miltä vastaaminen yleensä ottaen tuntui, oliko se hankalaa vai helppoa?

2. Miten vastaaminen eroaa vanhaan järjestelmään verrattuna?

3. Ottaisitko käyttöön sovelluksenlaisen Kysy kirjastonhoitajalta -palvelun, jos se olisi mahdollista, ottaen huomioon sovelluksen keskeneräisyyden?

4. Minkälaisia kehitysideoita sinulla on sovellukseen liittyen?

5. Minkälaisia kehitysideoita sinulla on loppukäyttäjän puoleen liittyen?