

Metropolia Ammattikorkeakoulu
Mediatekniikan koulutusohjelma

Mikko Sjöberg

**Yrityksen tietämyksenhallinnan kehittäminen tiedonhakua
parantamalla**

Insinööriyö 17.02.2010

Ohjaaja: IT-päällikkö Jouko Kanerva
Ohjaava opettaja: yliopettaja Kari Aaltonen

Tekijä Otsikko Sivumäärä Aika	Mikko Sjöberg Yrityksen tietämyksenhallinnan kehittäminen tiedonhakua parantamalla 79 sivua 17.02.2010
Koulutusohjelma	mediatekniikka
Tutkinto	insinööri (AMK)
Ohjaaja Ohjaava opettaja	IT-päällikkö Jouko Kanerva yliopettaja Kari Aaltonen
<p>Insinööriyössä tutkittiin rakennusalan tietopalveluyrityksen intranetin hakutoiminnon kehittämistä. Näkökulmana oli tietämyksenhallinta yrityksen liiketoimintaa edistävänä tekijänä. Yrityksessä tuotettujen dokumenttien ja verkkosivujen lisäksi myös sisäinen viestintä kuului tämän työn kannalta tietämyksenhallintaan. Olennaisen tiedon löytämisen helppous ja nopeus olivat kehitystyön arvioinnin kriteerejä. Keskeisenä tavoitteena oli tiedonhakua tehostamalla vapauttaa työaikaa tuottavampaan työhön.</p> <p>Kohteena ollut järjestelmä oli huhtikuussa 2009 käyttöön otettu intranet. Sen toimituksessa mukana tullut hakukone otettiin käyttöön vakioasetuksin, mikä asetti kehitysvaatimuksia hakutoiminnolle. Työ toteutettiin mahdollisimman pitkälle järjestelmän omia ominaisuuksia hyödyntäen, eikä varsinaista sovelluskehitystä tehty.</p> <p>Tutkimus pohjautui käyttäjätettiin, jolla selvitettiin todellisten loppukäyttäjien avulla hakukoneen käytettävyyttä. Käytettävyyden arvioinnissa mitattiin relevantin tiedon löydettävyyttä hakukonetta käyttämällä sekä tiedonhakuprosessin nopeutta. Lisäksi käyttäjillä oli mahdollisuus antaa vapaamuotoista palautetta. Tuloksissa löydettiin useita ongelmakohtia. Haluttu dokumentti ei hyvin valitulla hakusanalla listautunut riittävän ylös hakutuloksissa. Tulosten esittämisessä käytetyt metatiedot olivat sopimattomia, eikä hakukone osannut käsitellä suomen kielen yhdyssanoja oikein. Lisäksi haettava sisältöalue oli liian suppea ja käyttäjät toivoivat hakujen kohdentamismahdollisuuksia.</p> <p>Ratkaisuna yritykselle kehitettiin metatietomalli ja hakukonetta konfiguroitiin toimimaan paremmin loppukäyttäjien tarpeen mukaan. Haettavan tiedon relevanssia ja nopeutta voitiin parantaa merkittävästi. Samalla kehitettiin yrityksen sisäistä dokumentinhallintaa metatiedon hyödyntämisen avulla, mikä kehittää tiedonhakua paitsi hakukoneen avulla, myös perinteisen navigoinnin kautta. Tiedonhaun kehittäminen todettiin loppukäyttäjien keskuudessa tärkeäksi, ja tehdyt parannukset vastasivat hyvin käyttäjien tarpeita ja toiveita. Metadatan tehokas hyödyntäminen vaatii käyttäjiltä kuitenkin sitoutumista ja opettelua jokapäiväisissä työtehtävissä.</p>	
Hakusanat	tiedonhaku, tietämyksenhallinta, intranet, hakukone, tietotyö, metadatan, metatietomalli, tieto, informaatio

Author	Mikko Sjöberg
Title	Enhancing the knowledge management of a company by improving information retrieval
Number of Pages	79
Date	17 February 2010
Degree Programme	Media Technology
Degree	Bachelor of Engineering
Instructor Supervisor	Jouko Kanerva, IT-Manager Kari Aaltonen, Principal Lecturer
<p>The purpose of this thesis was to examine and improve a search function of a building information service company's intranet. The study focuses on knowledge management and how it can enhance the business activity of a company. The ease and quickness of retrieving relevant information were the main aspects to be evaluated. As a result, the company should be able to unleash employee's effective working time to more important tasks instead of searching information.</p> <p>The system studied was an intranet service that was deployed during April 2009. A search engine in a basic mode was included in the implementation. However, it was clear from the beginning that the search function does not meet the company's standards and needs for knowledge management. A configuration of the system using its own components was the main method in this thesis and no actual software development or design was made.</p> <p>The research was accomplished as a user test and the results were used in the search engine improvement. The real end-users were used as test persons to measure out how well the search engine's operation matched to the user needs. The main criteria were if the sought information was found at all and how high it was ranked in the result listing. Also, the time spent in the process was considered when drawing the conclusions. The end-users were also asked to give any free feedback of the search engine's operation.</p> <p>As a result, several demands for improvement were found. The wanted documents did not rank high enough in the search results list. The metadata used in representing the documents were irrelevant and inadequate. There were no special scopes for focusing the search and the compound words used in the Finnish language were processed incorrectly. Also, the content area of the search index was too limited.</p> <p>The most important improvements made and propositions recommended consisted of creating a metadata model to the company, increasing the relevance of the retrieved information and scoping a wider searchable content area.</p>	
Keywords	information retrieval, knowledge management, intranet, search engine, metadata, metadata model, knowledge work, information science

Sisällys

Tiivistelmä

Abstract

Lyhenteet, käsitteet ja määritelmät

1	Johdanto	6
2	Tiedon haku	7
2.1	Tiedon löytäminen tietämyksenhallinnan perustana	7
2.2	Yrityksen sisäinen verkkopalvelu	10
2.3	Verkkopalvelu tietovarastona	13
2.4	Hakumenetelmät	14
2.5	Hakupalvelun toimintaperiaate	17
2.6	Indeksi hakupalvelun sanavarastona	20
2.7	Hakutulosten esittämistapa tiedonhaun ohjaajana	23
3	Tiedonhakuprosessin laadun kehittämisen menetelmiä	29
3.1	Tiedonhaun kehittäminen edistää tietämyksenhallintaa	29
3.2	Semanttinen web tiedonhaun kehittäjänä	30
3.3	Ontologian käyttö sisällön kuvaamisessa	31
3.4	Metatiedon hyödyntäminen dokumenttien kuvailussa	33
3.5	Semanttisen webin sisällön mallintaminen	38
3.6	OWL verkkosisältöjen kuvailukielenä	39
4	Hakutoiminnon kehittäminen käyttäjätestauksen avulla	42
4.1	Tutkimuksen testimetodit ja arviointikriteerit	42
4.2	Testimalli	43
4.3	Tulosten arviointi	45
4.4	Heuristiset menetelmät hakutoiminnon kehittämisessä	46
5	Hakukonfiguraatio ja metatietomalli	47
5.1	Kehityksen kohteena ollut intranet	47
5.2	Käyttäjätestauksen tulokset	51
5.3	Käytetyn hakupalvelun sisäinen arkkitehtuuri	55
5.4	Tehdyt toimenpiteet	57
5.5	Ehdotus metatietomalliksi	62
5.6	Tehtyjen muutosten arviointi	66
6	Yhteenveto	68
	Lähteet	71
	Liitteet	
	Liite 1: Kyselylomake	74

Lyhenteet, käsitteet ja määritelmät

CMS (content management system): Sisällönhallintajärjestelmä, jota käytetään organisaatiossa hyödynnettävien dokumenttien ja muun sisällön hallintaan. Yleinen tyyppi on www-sivujen luomiseen ja muokkaamiseen käytetty tietojärjestelmä, jota hyödyntämällä käyttäjä ei tarvitse erikoisosaamista ja -ohjelmistoja sisällön ylläpitoon.

Domain-tunnus: Internetin verkkotunnus, jonka avulla numeroihin perustuvat IP-osoitteet voidaan korvata helpommin muistettavasti sanoin ja kirjainyhdistelmin.

Emulointi: Tietotekniikassa ohjelman tai laitteiston käyttäminen jossain muussa käyttöjärjestelmässä, kuin mihin se on alun perin tarkoitettu. Se tapahtuu emulaattorin avulla, joka on tätä käyttötarkoitusta varten suunniteltu tietokoneohjelma.

PDF (portable document format): Adobe Systems Inc:n erityisesti julkaisutoimintaa varten kehittämä käyttöjärjestelmästä riippumaton tiedostomuoto, jonka tarkoituksena on säilyttää dokumentin alkuperäinen muoto siirrettäessä sitä järjestelmästä toiseen. Julkaisujen lisäksi tiedostomuotoa käytetään laajasti sähköisessä dokumentinhallinnassa.

Skeema: Tietojenkäsittelyssä skeemalla tarkoitetaan yleisesti mallia, jolla kuvataan jotain järjestelmää abstraktilla tasolla. Sitä käytetään tiedon rakenteen ja sisällön kuvaamiseen sekä tietomallin käsitteistöjen välisten suhteiden kuvaamiseen.

Syntaksi: Luonnollisen tai formaalin kielen sisältämien merkkien yhdistelmien tarkasteluun käytetty lauseoppi.

Taksonomia: Alun perin biologiaa varten kehitetty luokittelujärjestelmä, jonka periaate on otettu käyttöön laajasti muilla aloilla. Taksonomiassa asiat tai käsitteet pyritään järjestämään luokkiin, joilla on alaluokkia. Luokkien väliset suhteet pystytään kuvaamaan systemaattisesti, ja alemmat luokat perivät ylemmän ominaisuudet.

URI (uniform resource identifier): Merkkijono, jota käytetään yksilöimään Internetissä oleva verkkoresurssi. Yksilöinnin avulla resursseja voidaan hyödyntää verkon kautta käyttämällä sopivaa tiedonsiirtoprotokollaa. Yleisin URI:n tyyppi on tavallinen www-osoite (URL).

W3C (World Wide Web Consortium): Maailmanlaajuinen konsortio, jonka tavoitteena on standardoida nykyisin käytössä olevia web-teknologioita sekä kehittää kokonaan uusia ratkaisuja tulevaisuudessa.

1 Johdanto

Insinööriyön tavoitteena oli kehittää yrityksen intranetin hakutoimintoa osana tietämyksenhallintaa. Yrityksen tietämyksenhallinnalla tarkoitetaan tässä työssä sellaisia toimenpiteitä ja prosesseja, joilla yritys pyrkii hallinnoimaan systemaattisesti täsmällisen tiedon lisäksi sen käytettävissä olevaa hiljaista tietoa. Perinteisen dokumentinhallinnan lisäksi tietoa kertyy ja kerrytetään runsaasti sisäisten viestintäkanavien kautta. Tieto saadaan keskitetysti yrityksen työntekijöiden käyttöön intranetin avulla. Hakutoiminnon tehtävä on toimia tietämyksenhallinnan tehostajana, joka auttaa työntekijää hyödyntämään nopeasti ja tehokkaasti eri työprosesseissa tarvittavaa tietoa. Tiedonhakuprosessia kehittämällä voidaan vapauttaa työntekijöiden aikaa varsinaiseen tuottavaan työhön.

Tässä insinööriyössä arvioitiin haun kehittämistä yhden järjestelmän ominaisuuksien perusteella. Kehittämällä ei tässä tarkoiteta varsinaista sovelluskehitystä, vaan järjestelmän tukemien ominaispiirteiden ja niiden konfigurointimahdollisuuksien tutkimusta. Intranetin muiden ominaisuuksien tai toiminnallisuuden kehittämiseen tai niiden arviointiin ei oteta kantaa.

Insinööriyön tilaajana oli Rakennustieto Oy (myöhemmin *tilaaja*), joka on rakennusalaan erikoistunut kustannus- ja tietopalveluyritys. Yrityksessä työskentelee noin 90 työntekijää, sen liikevaihto on noin 11 miljoonaa euroa (2007) ja sillä on toimintaa Suomen lisäksi Venäjällä ja Virossa. Tilaaja on kehittänyt voimakkaasti viime vuosina tietopalvelutuotteidensa digitalisointia ja siirtynyt pois painotuotteista verkkopalveluiden tai muiden sähköisten medioiden kautta jaettaviin tai käytettäviin tuotteisiin.

Työn lähtökohtana oli tilaajan sisäinen kehitysprojekti, jonka tarkoituksena on parantaa tilaajan tietämyksenhallintaa työprosesseja, sisäistä viestintää ja dokumenttienhallintaa kehittämällä. Projektin keskeisin ja tilaajan loppukäyttäjille näkyvin tulos on uudistettu intranet, joka otettiin käyttöön huhtikuussa 2009. Uudistetussa intranetissä oli ensi kertaa koko henkilöstön käytössä sanahakuun perustuva hakukone, jonka käytettävyyden ja ominaisuuksien laadun kehittäminen ovat tämän insinööriyön kohteena. Työn arvioinnissa keskeistä on pyrkiä selvittämään, miten hyvin hakutuloksena esitetyt dokumentit vastaavat loppukäyttäjän tarpeita.

2 Tiedon haku

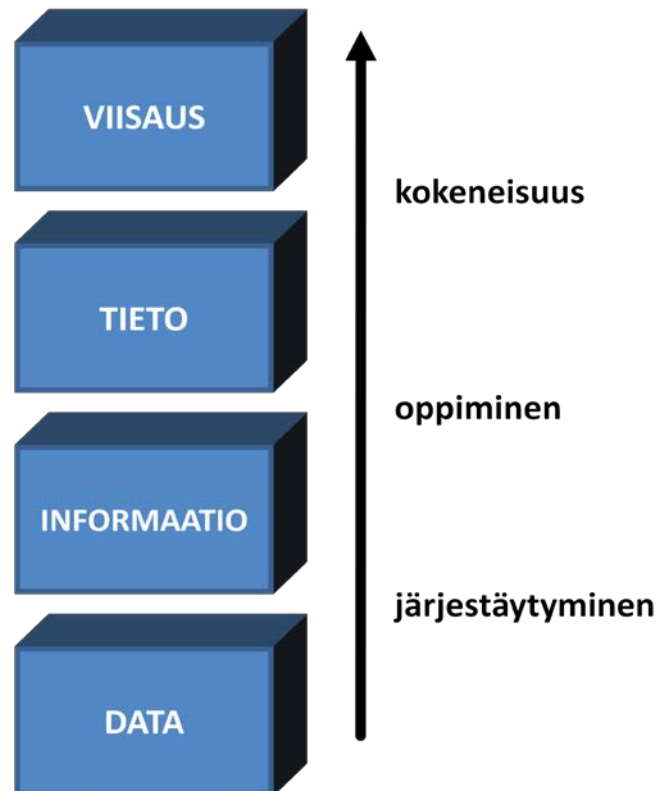
2.1 Tiedon löytäminen tietämyksenhallinnan perustana

Tietotyön yleistymisen myötä tiedosta ja sen tehokkaasta hyödyntämisestä on tullut yrityksille keskeinen menestystekijä [1, s. 4]. Kokonaisuutta kutsutaan suomeksi yleisesti tietämyksenhallinnaksi (engl. knowledge management) erotuksena tiedonhallinnasta, joka on terminologiassa liitetty yleisesti tietokantojen suunnitteluun ja ylläpitoon [2, s. 48-50]. Lisäksi knowledge management on hieman näkökulmasta riippuen suomennettu myös tietojohdantamiseksi tai osaamisen johtamiseksi [3, s. 117]. Tietämyksenhallinta on aiemmin ollut suuryritysten liiketoiminnan kehittämisen väline. 2000-luvulla myös pk-yritykset ovat alkaneet hyödyntää sitä omassa toiminnassaan. [1, s. 5] Tietämyksenhallinnan lähtökohtana on, että yrityksen työntekijöiden luovuus rakentuu yhä enemmän käytettävissä olevan, systemaattisesti hallitun tietopääoman varaan [4, s. 109–110].

Tiedon määrän kasvu asettaa haasteita tiedon löydettävyydelle ja merkitykselle ja sitä kautta myös sen hyödynnettävyydelle. Tietotyön tuottavuuden kannalta olennaista on aiemmin tuotetun tiedon mahdollisimman helppo saavutettavuus. Tutkimustulosten mukaan työntekijät saattavat käyttää työajastaan jopa 15–35 % tiedonhakuun. Hakuprosessissa aikaa käytetään paitsi yksittäisen tiedon löytämiseen, myös sisällön merkityksen arviointiin vertailemalla eri tietolähteitä keskenään. [5]

Tietotyössä on olennaista erottaa merkityksellinen tieto erilaisten lähteiden tarjoamasta informaatiosta. Tietotekniikassa termejä ”data”, ”informaatio” ja ”tieto” on käytetty merkityksiltään väljästi ja usein päällekkäisesti [6, s. 29, 48]. Informaatio on vastaanottajalleen merkityksen tarjoava joukko dataa, joka on jäsenetty jonkin järjestelmän mukaan. Esimerkiksi suomen kielen yksittäinen sana koostuu aakkosista ja kieliopin mukaisesti toisiinsa liitetyt sanat muodostavat lauseen. Sillä on vastaanottajalle merkitys, mikäli hän ymmärtää käytetyn järjestelmän tavan koodata datasta informaatiota. Lauseen sisältö muuttuu tiedoksi vasta, jos vastaanottaja osaa tulkita merkityksen ja muuttaa toimintaansa tai ajatteluaan merkityksen mukaan [4, s. 115–116].

Kuvassa 1 on mallinnettuna tiedon käsitteen rakentumista alhaalta ylöspäin rakentuvana hierarkiana.



Kuva 1. Tiedon hierarkia [7, s. 6]

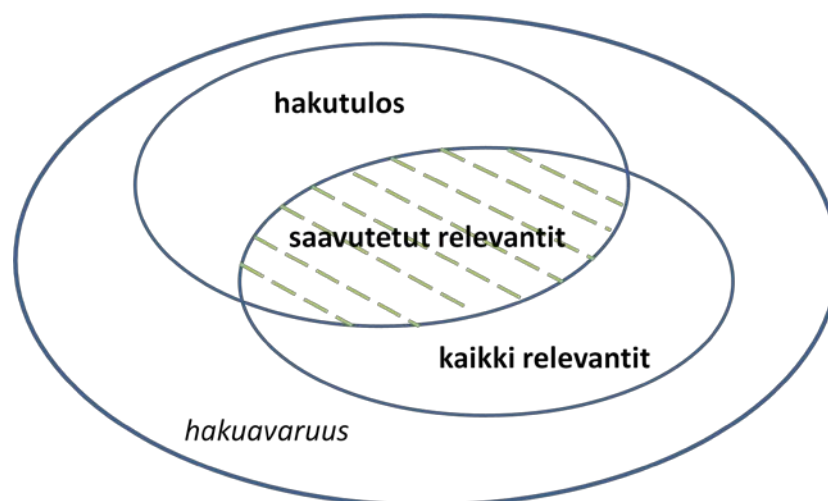
Hierarkian osien suhteet esitetään tässä siten, että datalla ei ole mitään kontekstia, toisin sanoen se ei sellaisenaan ole liitettävissä mihinkään. Kun data voidaan esittää jossain kontekstissa, siitä tulee informaatiota. Tiedosta voidaan puhua siinä vaiheessa, kun informaation palasia opitaan yhdistämään ja muodostamaan kokonaiskäsitteeksi asiasta. Vastaavasti, kun tiedon syntymisen taustalla oleva mekanismi opitaan tiedostamaan ja hyödyntämään myös toisissa aihealueissa, voidaan puhua viisaudesta. Tietämyksenhallinnan keskeisenä visiona on kehittää yritystä kohti viisautta. Käytännön toiminnan tasolla voidaan erottaa kaksi tietämyksenhallinnan koulukuntaa: tiedon tuotantoprosessin kehittämiseen ja tiedon hallintaan ja mittaamiseen keskittynyt. [7, s. 6–7]

Tiedonhaun keskeinen ongelma on erottaa haettavaan asiaan olennaisesti liittyvä tieto sitä lähellä olevasta tai siihen kokonaan liittymättömästä informaatiosta. Tätä kutsutaan yleisesti tiedon relevanssiksi. Se on keskeinen käsite ja mittari tiedonhakuprosessin onnistumisen arvioinnissa tiedonhakujärjestelmästä riippumatta. Tiedonhakijan merkityk-

selliseksi kokema informaatio on *relevanttia* ja vastaavasti muu kohinan sekaan jäävä *irrelevanttia*. [8, s. 79]

Relevanssin mittaamiseen on kehitetty erilaisia menetelmiä, joista yleisimmin käytettyjä ovat Term Frequency (TF) ja Inverse Document Frequency (IDF). Term Frequencyllä mitataan yksittäisen sanan esiintymistiheyttä suhteessa koko dokumentin laajuuteen. Inverse Document Frequencyllä mitataan relevanssia määrittämällä koko dokumenttiavaruuden suhde niihin dokumentteihin, joissa etsitty sana esiintyy. TF ja IDF ovat perustason mittaamenetelmiä. [8, s. 81–84] Relevanssin mittausta on pyritty kehittämään sisällön merkitystä korostavaan suuntaan ottamalla huomioon lauserakenteita ja sanojen esiintymistä suhteessa toisiin sanoihin [9]. Koska tiedon relevanssi on käyttäjälle henkilökohtainen asia ja tietosisällön merkitys vaihtelee kontekstin mukaan, ei hakupalvelua voi kehittää yksiselitteisesti toimivaksi. Haun jatkuva kehittäminen käyttäjäpaltteen keräyksen ja tutkimuksen avulla onkin tärkeää. [8, s. 79]

Kuvassa 2 on esitetty relevanttien hakutulosten suhdetta kaikkiin dokumentteihin, jotka ovat sisällytettynä koko *hakuavaruuteen*. Hakuavaruudella tarkoitetaan kaikkia niitä dokumentteja, jotka on sisällytetty haun piiriin. Saavutetut relevantit hakutulokset ovat leikkaus kaikkien hakutulokseen listattujen ja kaikkien hakuavaruudessa esiintyvien relevanttien dokumenttien välillä. [8, s. 25–26]



Kuva 2. Haussa saadut relevantit tulokset suhteessa muihin [8, s. 26]

Hakupalvelun toiminnan *tarkkuudella* (precision) tarkoitetaan järjestelmän kykyä erottaa hakutulokseen vain relevantteja dokumentteja. Vastaavasti *saanti* (recall) on järjestelmän kyky sisällyttää mahdollisimman suuri osa koko hakuavaruuden relevanteista dokumenteista hakutulokseen. [10, s. 110–111]

Keskeisenä ongelmana sanahaun hyödyntämisessä on sanojen monimerkityksellisyys: samalla sanalla voi olla useita eri merkityksiä ja vastaavasti samaan käsitteeseen voidaan viitata useilla eri sanoilla. Tästä on seurauksena, että tarvittava tieto saattaa hävitä kohinaan (semantic noise). Tiedonhakuprosessin kehittämisessä on olennaista pystyä määrittelemään haun kohteena oleva tieto niin, että se on merkitykseltään mahdollisimman yksiselitteistä [11].

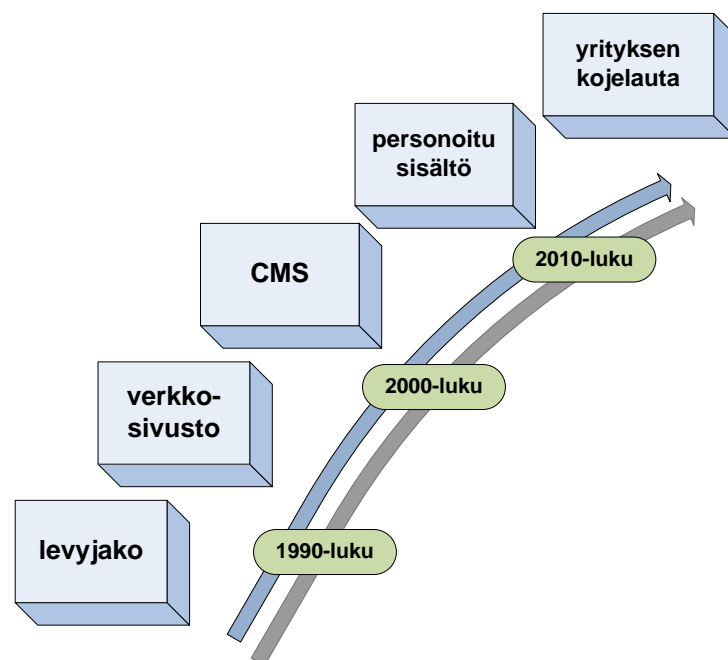
2.2 Yrityksen sisäinen verkkopalvelu

Intranetillä tarkoitetaan pelkistetyssä muodossa TCP/IP-yhteyskäytäntöjä käyttävää organisaation sisäistä tietoverkkoa [12]. Käytännössä nykyaikaisella intranetillä tarkoitetaan www-selaimella käytettävää yrityksen sisäverkossa toimivaa verkkopalvelua, johon on pääsy ainoastaan yrityksen tai organisaation henkilöstöllä tai jäsenellä. Käytön rajaus sisäverkkoon ja tarkasti rajattuun käyttäjäkuntaan erottaa sen Internetistä ja ulkopuolisten sidosryhmien käyttöön tarkoitetuista ekstranet-palveluista. [7, s. 21] Isoilla organisaatioilla intranet saattaa olla jaettuna myös useisiin pienempiin intraneteihin alueellisesti tai toimialoittain. Sisällöltään laajaa ja toiminnallisuuksiltaan monipuolista intranetiä voidaan kutsua myös yritysportaaliksi. [13, s. xiii]

Nykyaikaisella intranetillä voi olla hyvin monia käyttötarkoituksia. Alkuvaiheessa 1990-luvulla ensimmäiset ratkaisut olivat lähinnä jaettuja tiedostopalvelimia, joissa oli jokin ennalta sovittu, yrityksen sisäistä rakennetta ja toimintaa tukeva kansiorakenne. Rakennetta käytettiin yrityksen sisäisesti käytettävien yhteisten dokumenttien jakamiseen. Tästä kehittyneempi versio oli alkeellinen verkkosivusto, joka oli vielä hankalasti ylläpidettävissä ja vaati erityisosaamista. Seuraavassa vaiheessa intranet kehittyi viestintäkanavaksi, jossa pystyttiin keskitetysti viestimään organisaation jäsenille ja sisällön ylläpito ei vaatinut erityisosaamista sisällönhallintajärjestelmien (content management system, CMS) käyttöönoton ansiosta. Hajautetun ylläpidon avulla periaatteessa kuka

tahansa saattoi tuottaa sisältöä. Tämä aiheutti kuitenkin sisällön määrän ja rakenteen hallitsemattomuutta. [3, s. 23]

Intraneteissä pyritään siirtymään tietämyksenhallintaa paremmin soveltaviin ratkaisuihin. Tämä tarkoittaa käytännössä sisällön personointia ja toteuttamista siten, että se tukee mahdollisimman hyvin yksittäisiä työprosesseja ja työntekijän niissä tarvitsemaa tietoa ja työmenetelmiä. Kuvassa 3 on esitetty intranetin kehityksen eri ajanjaksoissa erotettavissa olevia vaiheita.



Kuva 3. Intranetin kehitysvaiheita [3, s. 23–24]

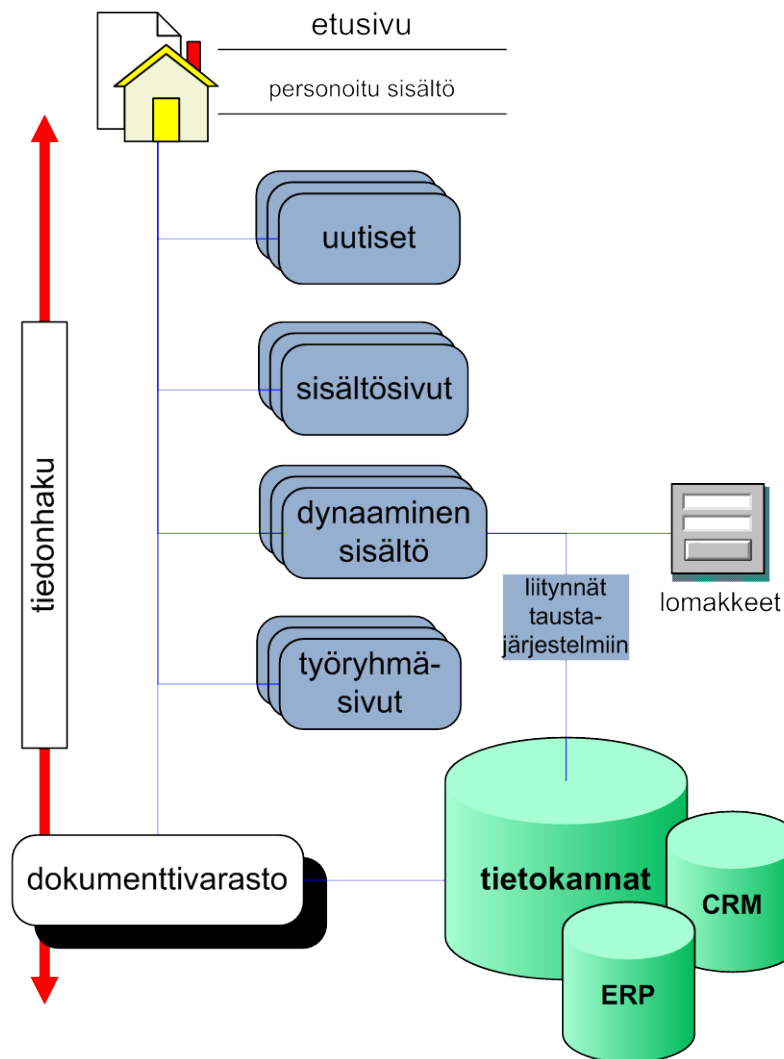
Kehittyneimmällä tasolla intranetiä on alettu kutsua yrityksen sisäiseksi portaaliksi tai sisäiseksi kojelaudaksi [3, s. 24]. Portaali esitetään tässä verkkosivustona, johon on koottu tietyn rajatun käyttäjäryhmän (esim. yrityksen henkilöstö) kannalta yhteistä mielenkiintoa herättävää informaatiota [14, s. 40–41]. Portaalien sisältö ja palveluarkkitehtuuri voivat olla hyvin monipuolisia. Sisältö voi koostua lukuisista erilaisista tietolähteistä. Tyypillisesti intranetiin tuodaan yrityksen taloudellisesta tilasta kertovaa informaatiota erilaisilla raportointityökaluilla, joita käytetään yrityksen toiminnan ohjaukseen. Lisäksi intranet voi toimia käyttöliittymänä useisiin taustajärjestelmiin, kuten yrityksen toiminnanohjausjärjestelmä, mikä on tuonut johtamisjärjestelmään liittyviä piir-

teitä intranetin käyttöön. [3, s. 43] Vastaavasti sellaiset Web 2.0 -käsitteistön sosiaaliseen mediaan liittyvät ominaisuudet kuin blogi, wiki ja vapaamuotoinen keskustelu ja ryhmien muodostaminen ovat tehneet intranetistä yhteisöllisen kohtaamispaikan, jota voidaan hyödyntää yrityksen sisäisen ilmapiirin kehittämisen ja motivaation kasvattamisen välineinä. [15, s. 18–19]

Yrityksen intranetin tai sisäisen portaalin avulla tavoiteltaviksi hyödyiksi voidaan tiivistää

- yhtenäinen kokonaiskuva organisaatiosta
- tiedon järjestäminen ja hakumahdollisuus
- välitön mahdollisuus hyödyntää yrityksen tietopääomaa ja resursseja
- suora linkitysmahdollisuus liiketoimintatietoon
- hiljaisen tiedon ja osaamisen saatavuus ja hyödynnettävyys
- sisällön personointi [13, s. 28].

Kuvassa 4 on nykyaikaisen intranetin rakennekaavio, jossa on esitettyä myös liitännät taustajärjestelmiin.



Kuva 4. Nykyaikaisen intranetin rakenne

Kuva perustuu pääosin tämän insinööriyön kohteena olleen intranetin rakenteeseen siltä osin, kuin se on toteutettuna nyt ja suunniteltu kehitettäväksi jatkossa. Tiedonhaku on kuvattu koko palveluun kohdistuvana, sisältöä integroivana tekijänä.

2.3 Verkkopalvelu tietovarastona

Internetin tärkeimmäksi käyttömuodoksi on muodostunut tiedon haku käyttämällä hyväksi erilaisia hakukoneita (esim. Google). Käytännössä tämä tarkoittaa sitä, että Internet toimii laajana tietovarastona, jossa on hajautettuna hyvin suuri määrä erityyppistä ja -tasoista informaatiota. Tiedon luotettavuus on usein kyseenalaista ja tiedonhakijalta vaaditaan korkeaa lähdekriittisyyttä ja medialukutaitoa. Tietosisällöt muuttuvat usein

nopeasti ja saattavat vaihtaa paikkaa tai lakata olemasta. Tiedon heikko strukturointi on johtanut siihen, että sanahakuun perustuvat hakukoneet ovat muodostuneet välttämättömiksi verkkopalvelun käytettävyyden kannalta. [16, s. 3]

Koko Internetin kattamaan pyrkivien hakupalveluiden lisäksi myös yksittäisille sivustoille tehdään nykyisin omat hakupalvelunsa. Toimiva hakupalvelu on erittäin tärkeä tekijä myös, kun arvioidaan yksittäisen sivuston käytettävyyttä. Mikäli loppukäyttäjä ei löydä sivustolta riittävän nopeasti haluamaansa tietoa, hän siirtyy helposti käyttämään toista tietolähdettä [17]. Käyttäjä haluaa ainoastaan relevantteja dokumentteja. Perinteisen hakupalvelun perusongelma on, että se vaatii käyttäjän omaa päättelyä ja valintaa useiden rinnakkaisten, mutta sisällöltään eriävien hakutulosten joukosta, ennen kuin relevantti tieto on hyödynnettävissä. [18, s. 116] Intranet ja sen hakutoiminto on teknisesti rinnastettavissa yksittäiseen Internet-sivustoon, vaikka sen käyttäjäryhmä on tarkkaan rajattu.

Verkkopalvelun hyödynnettävyys tietovarastona korostuu tiedon yhdistely- eli integrointimahdollisuuden kautta. Yksi relevantti hakutulos johtaa usein tiedon täydentämiseen uudella haulla. Lopputuloksena käyttäjä muodostaa tuloksista kokonaisuuden yhdistelemällä eli *integroimalla* informaatiota eri lähteistä tiedoksi. Integraatio merkitsee hakuprosessin kykyä kerätä, yhdistää ja esittää mahdollisimman tehokkaasti kaikki se olennainen tieto, joka palvelee käyttäjää senhetkisessä tarpeessa. Tiedon yhdistelyssä on pyritty automaatioon, jossa yksinkertaisimmat haut voisivat tuottaa käyttäjän oletettavasti tarvitsemaa lisätietoa. Tällainen mahdollisuus on esimerkiksi kirjaan kohdistettu haku, joka hyödyntää muita verkkopalveluita Web Services -teknologian avulla ja voi tuottaa teoksen bibliografisen tiedon lisäksi vertailevaa hintatietoa eri myyntipaikoista. [16, s. 4]

2.4 Hakumenetelmät

Verkkopalvelun hyödyntäminen tietolähteenä ja -varastona perustuu Internetin käyttötottumuksien muokkaamiin toimintamalleihin ja -tapoihin. Vaikka yrityksen sisäinen verkkopalvelu intranet onkin sisällöltään käytännössä aina rajattu, se voi sisältää hyvin suuren määrän informaatiota, jonka jäsentämiseksi ja tavoittamiseksi täytyy olla syste-

maattinen tapa etsiä tietoa. Verkkopalvelun tiedonetsintämenetelmissä voidaankin perustasolla erottaa kolme eri lähestymistapaa: suora navigointi sivuston osoitteen avulla, hakemistopohjainen navigointi tai hakukoneen käyttö [8, s. 19].

Tiedonhaku suoraan sivuston osoitetta hyödyntämällä

Yksinkertaisin tapa hakea tietoa verkosta on kirjoittaa sivuston osoite suoraan selaimen osoiteriville ja siirtyä suoraan palvelun verkkosivulle. Tämä toimii tyypillisesti tilanteessa, jossa kohde on yleisesti tunnettu ja käyttäjä joko tietää entuudestaan tai voi riittävän tarkasti päätellä osoitteen (esim. www.nokia.com). Tällöin käyttäjä tietää verkkoosoitteen muodostamistavan ja osaa päätellä osoitteen yrityksen nimen ja domain-tunnuksen perusteella. Sivustolle siirtyminen tapahtuu näin mahdollisimman suoraan. Tiedonhaun näkökulmasta käyttäjä ei tässä tilanteessa välttämättä tiedä tarkalleen, mitä tietoa haluaa, mutta olettaa tiedon löytyvän esimerkiksi laitteen valmistajan kotisivuilta. Toinen vaihtoehto suoran navigoinnin hyödyntämiseen on tilanteessa, jossa käyttäjä haluaa selailta yleisesti esimerkiksi yrityksen sivuja kokonaiskuvan hahmottamiseksi. [8, s. 19–20]

Hakemistorakenteeseen tai portaaliin perustuva tiedonhaku

Intranet varsinkin pitkälle kehitettynä voidaan nähdä yrityksen sisäisenä portaalina, joka toimii väylänä kaikkeen yrityksen sisäiseen tietoon. Portaaliksi käsitetään Internetissä sellainen verkkopalvelu, joka toimii aloituspisteenä selailulle ja joka ei tarjoa itsenäisenä palveluna kohdennetusti tuotettua tietoa, vaan on linkkikokoelma useiden eri tiedontuottajien palveluiden sisältöön. Tällaisia palveluita ovat muun muassa msn.com ja suomi.fi.

Tiedonhaun kannalta olennaista portaalissa on, että sisältö on luokiteltu tai jäsennetty jollain systemaattisella ja loogisella tavalla hakemistoiksi. Tyypillisesti sivusto on jaettu alihakemistoihin aihepiireittäin, esimerkiksi uutiset - viihde - urheilu - matkailu. Lisäksi sivustolla voi olla asiasanahakemisto tai muu systemaattinen luokittelu tukemassa tiedonhakuprosessia. Käyttäjä etsii tietoa selailemalla häntä kiinnostavaa aihepiiriä käsittelevän hakemiston sisältöä. Sisällöstä on runsaasti linkkejä aihepiiriin liittyviin sivustoihin. Käyttäjä etenee linkkien avulla kohti haluamaansa sisältöä tai sitten palaa takaisin

ylemmille tasoille ja valitsee jonkin toisen, paremmin soveltuvan polun tiedonhakuprosessilleen. [8, s. 20–21]

Intranet-palvelun sivustorakenne toimii vastaavasti yrityksen sisäisenä portaalina ja tiedon systemaattisena rakenteellisena esittämistapana. Rakenne voi pohjautua esimerkiksi organisaatorakenteeseen tai tuoteryhmiin tai näiden yhdistelmään. Olennaista hakemistopohjaisessa tiedonhaussa on kuitenkin, että se palvelee parhaiten käyttäjää, joka ei ole kovin hyvin selvillä siitä, mitä on hakemassa. Tällöin aihepiirin hahmottaminen kokonaisuutena voi auttaa käyttäjää löytämään yksittäisen tiedon. Ensisijaisesti tämäntyyppinen, selaileva tiedonhaku kuitenkin auttaa kokonaisuuksien oppimisessa ja hahmottamisessa.

Tiedonhaku hakukoneella

Hakukoneiden teknisen kehittymisen myötä ne ovat nousseet tärkeimmiksi verkkopalveluiden tiedonhakumenetelmiksi. Hakukoneella tapahtuva tiedonhakuprosessi voidaan tiivistää neljään vaiheeseen:

a) Kyselyn muodostaminen

Käyttäjä kirjoittaa hakukoneen syötekenttään yhden tai useamman *hakusanan* (engl. keyword). Merkkijono lähetetään hakukoneelle käsiteltäväksi. Käyttäjä voi tehostaa hakusanojen käsittelyä *loogisilla operandeilla* (ks. luku 2.5).

b) Valinta tuloksesta

Hakukone palauttaa hakutuloksen, joka on listaus niistä sivuista tai dokumenteista, jotka parhaiten vastaavat kyselyä hakukoneen sisäisen luokitusjärjestyksen (engl. ranking) mukaan. Otsikko toimii linkkinä sivulle tai dokumenttiin, ja käyttäjä siirtyy sitä kautta mielestään hakuaan parhaiten vastaavalle sivulle. Hakutulossivu voi sisältää erilaista metatietoa, joka helpottaa kohteen merkityksen arviointia.

c) Navigointi

Käyttäjä selailee avaamaansa sivua tai dokumenttia ja tutkii, löytyykö haluttu tieto siitä tai edelleen linkityksen kautta muualle tai alisivuille navi-

goimalla. Käyttäjä voi palata takaisin hakutulossivulle ja valita jonkin toisen tuloksen.

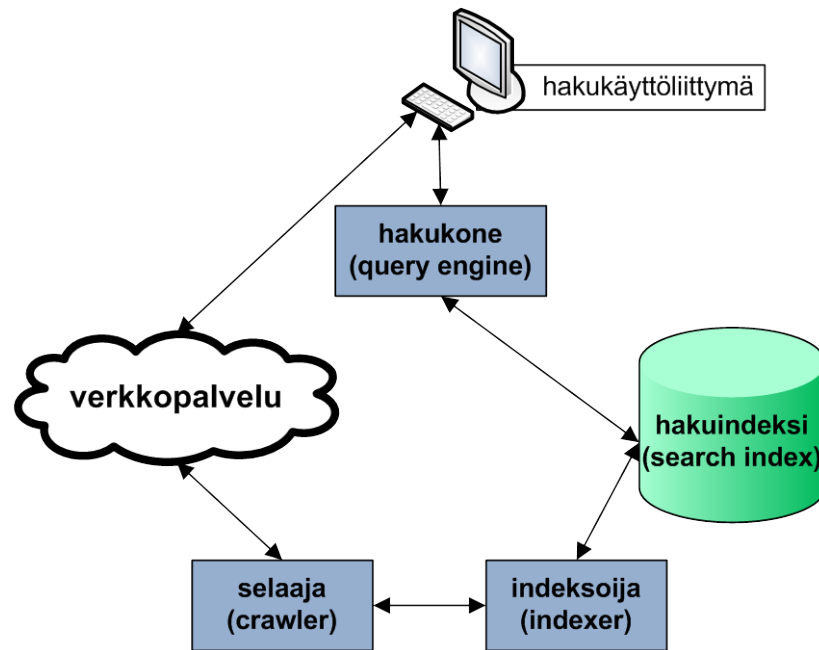
d) Kyselyn muokkaaminen

Mikäli hakutulos ei tuota tulosta, käyttäjä voi palata takaisin muokkaamaan kyselyä paremmaksi. Tällöin hakuprosessi alkaa uudelleen kohdasta a. [8, s. 21–22]

2.5 Hakupalvelun toimintaperiaate

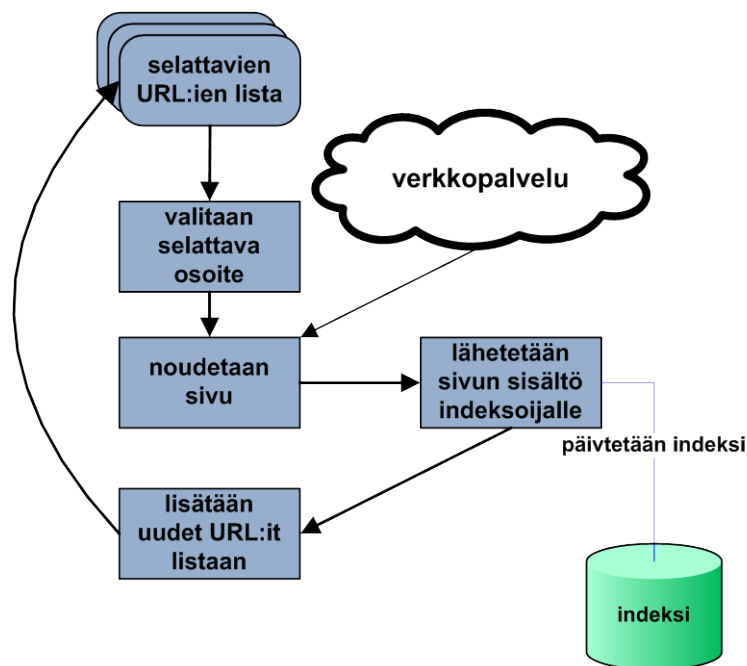
Verkkopalvelun hakutoiminto perustuu palvelimilla olevan sisällön keräämiseen, analysointiin ja järjestämiseen. Toimenpiteiden päämääränä on muokata sisältö sellaiseen muotoon, josta voidaan tuottaa käyttäjän kannalta mahdollisimman relevantteja hakutuloksia. Sisällön keruuta ja järjestämistä kutsutaan indeksoinniksi, joka muodostaa indeksitaulun. Keräys tehdään ohjelmallisesti selaamalla palvelu läpi ja rakentamalla indeksitaulu löydetyin sisällön mukaan. Indeksien rakentamisessa koko palvelun sisältö luetaan koneellisesti ja siihen liittyvä muu data, kuten kuvat, videot ja liitetiedostot, tallennetaan myös indeksiin. [16, s. 17–19]

Kuvassa 5 on yksinkertaistettu malli hakupalvelun rakenteesta. Sen tärkeimmät komponentit ovat *selaaja* (crawler), *indeksoija* (indexer), *hakuindeksi* (search index), *hakukone* (query engine) ja *käyttöliittymä* [8, s. 66].



Kuva 5. Hakupalvelun yksinkertaistettu arkkitehtuuri [lähteen 19 mukaan]

Hakupalvelun toiminta alkaa sisällön keruulla, joka tapahtuu selaamalla verkkopalvelun sisältö systemaattisesti läpi jostain selausalgoritmia käyttämällä. Käytännössä tämä tarkoittaa ohjelmistoa (web crawler), joka aloittaa selaamisen jostain sille määritellystä aloitusosoitteesta (seed URL). [16, s. 22] Laajoissa hakupalveluissa aloitusosoitteita on useita. Ohjelman avulla seurataan aloitussivulta lähteviä linkkejä ja ladataan linkitetyt sivut sisältöineen hakuindeksin muodostamista varten niin pitkälle, kuin linkitettyjä sivuja löytyy [19]. Kuvassa 6 on yksinkertaistettu malli selausalgoritmin toiminnasta.



Kuva 6. Yksinkertaistettu kuvaus indeksoijan toiminnasta [lähteen 8, s. 70 mukaan]

Selattavien osoitteiden määrä päivittyy näin ollen jatkuvasti, koska verkkopalvelun sisältö muuttuu. Algoritmia voidaan kehittyneemmällä tasolla muokata, ja se voidaan määrittää toimimaan tiettyjen ehtojen mukaisesti, esimerkiksi selaamaan indeksistä vain vanhentuneet sivut. Tämän jälkeen selaus lopetetaan, vaikka selattavia sivuja olisi muuten jäljellä. Selaamiseen voidaan sisällyttää myös erilaisia sääntöjä. Tällaisia ovat esimerkiksi tarve poistaa indeksistä sellaiset sivut, jotka ovat sisällöltään identtiset mutta joilla on eri osoite. Sisällön koostamisessa käytetään yleensä jonkinlaista suosituimmuusjärjestystä sivuille. Se perustuu yleisimmin siihen, kuinka monelta sivulta on viittauksia (linkityksiä) selattavalle sivulle. Algoritmin optimoiminen on haastavaa, koska verkkosivujen rakenne ei ole standardi, vaan sivut sisältävät nykyisin rakenteellisesti paljon dynaamisia osia tai ne saattavat olla kokonaan dynaamisesti palvelimen tietokannasta tuotettuja. Tämä tarkoittaa sitä, että sivun olemassaoloa on vaikea ennustaa, sisällöt voivat muuttua hyvin nopeasti ja sivu voi sisältää ohjelmallisia komponentteja, joiden sisältöä selaus ei pysty tavoittamaan. [8, s. 70–71]

Koko Internetin sisältöön hakumahdollisuuden tarjoavat hakupalvelut kohdistavat sisällön selaamisen vastaavasti mahdollisimman laajasti Internetiin. Tällöin aloituspisteinä toimivien juuriosoitteiden täytyy olla erityisen huolellisesti valittuja, jotta mahdolli-

simman suuri osa verkon sisällöstä saadaan selattua sisällön indeksointia varten. Selauksen yhteydessä löydetty sivut koostetaan indeksoijan avulla hakuindeksiksi, joka toimii koko hakupalvelun sisältämien hakusanojen talletuspaikkana. Hakuindeksistä on tarkemmin omassa luvussaan ”2.6 *Indeksi toimii hakupalvelun sanavarastona*”. [8, s. 67]

Hakupalvelun ytimenä toimii hakukone (query engine). Se yhdistää käyttöliittymän avulla hakuindeksin sisällön, käyttäjän ja verkkopalvelun sisällön hakupalveluksi, jota voi hyödyntää tiedon haussa. Hakukoneen toiminta on kaksivaiheinen. Ensin se vertailee käyttöliittymän kautta sille syötettyjä hakutermejä indeksin kanssa ja palauttaa niitä parhaiten vastaavien dokumenttien URL:t. Tämän jälkeen hakukone järjestää dokumentit järjestykseen relevanssin mukaan parhaiten kyselyä vastaava dokumentti ensimmäisenä. Hakukoneeseen kuuluu tyypillisesti myös mahdollisuus kohdentaa tai yksilöidä kyselyä tietyissä rajoissa. Tällaisia keinoja voivat olla esimerkiksi

- sanojen pakotettu poissulkeminen hakutulospöytäkirjasta
- sanojen pakotettu sisällyttäminen hakutulospöytäkirjaan
- pakotettu lausehaku käyttämällä lainausmerkkejä lauseen ympärillä
- sanan katkaiseminen käyttämällä niin sanottua jokerimerkkiä (yleensä asteriski (*) tai prosentti (%)), millä haetaan kaikkia niitä merkkijonokombinaatioita, jotka sisältävät hakusanan muut merkit (esim. kysely ”kirja*” tuottaa tällöin vastauksissa sanat kirjakauppa ja kirjahylly)
- Boolean operaattoreiden JA - TAI (AND - OR) hyödyntäminen hakutermin yhdistelyssä. Useimmat hakukoneet tekevät kyselyn käyttämällä välilyöntiä automaattisesti JA-operaattorina. [8, s. 69, 80]

Hakutulosten esittämisestä on tarkemmin luvussa ”2.7 *Hakutulosten esittämistapa ohjaa tiedonhakua*”.

2.6 Indeksi hakupalvelun sanavarastona

Hakuindeksi on keskeinen osa hakupalvelun arkkitehtuuria. Sen tärkein tehtävä on tiivistää verkkosivusto(je)n sisältämä informaatio sellaiseen muotoon, että siihen voidaan

kohdistaa mahdollisimman tehokkaita kyselyitä. Käyttäjän kannalta tämä merkitsee ajantasaisen ja merkittävän tiedon mahdollisimman nopeaa ja helppoa löydettävyyttä. [16, s. 17–18] Ilman indeksointia hakukone kävisi hakua tehdessään aina läpi koko aineiston kaikki dokumentit, mikä vaatisi niin paljon laskentatehoa ja aikaa, että haun suorittamien ei olisi mielekäästä.

Hakuindeksi perustuu tietorakenteeltaan käänteisen hakemiston periaatteeseen. Tämä on käytännössä hyvin pitkälle analoginen kirjojen aakkosellisen asiasanahakemiston kanssa, missä sisällön keskeiset termit ovat listattuina ja niihin on liitetty sivunumeroviittaukset. Vastaavasti verkkopalvelun indeksi koostetaan selatusta sisällöstä, jonka kaikki sanat ovat listattuina aakkosjärjestykseen ja niihin on liitetty viittaukset, millä sivuilla kukin sana esiintyy. Koko Internetin kattavan hakupalvelun indeksin sisältämien sanojen määrä lasketaan kymmenissä miljoonissa. [8, s. 67] Suomen kielessä on sanoja noin puoli miljoonaa, tosin lukumäärä vaihtelee sen mukaan, otetaanko murteet mukaan ja toisaalta kieli elää ja uudistuu koko ajan [20].

Hakuindeksiin tallennetaan paljon muutakin informaatiota indeksoidusta verkkopalvelusta. Tällaista tietoa voi olla esimerkiksi se,

- kuinka monessa dokumentissa yksittäinen sana esiintyy
- kuinka monta kertaa sama sana toistuu dokumentissa
- missä kohdin dokumenttia sana esiintyy. [8, s. 67]

Indeksi tallennetaan indeksitauluun, joka on käytännössä tietokanta, johon hakukone kohdistaa kyselyn. Taulussa esitetään systemaattisesti kaikki selauksessa kerätty tieto. Taulukossa 1 on yksinkertaistettu esimerkki indeksitauluun kerätystä tiedosta.

Taulukko 1. Indeksitaulun muodostus [16, s. 18]

Selauksessa löytyneet sanat	Selattu osoite	Sanan esiintymistiheys
Sana ₁	URL ₀	1–N kpl
Sana ₂	URL ₀	1–N kpl
⋮	⋮	1–N kpl
Sana _N	URL ₀	1–N kpl

Taulukon esimerkissä on selattu jokin verkko-osoite URL₀, josta selaaja on löytänyt N kpl sanoja. Kukin sana esiintyy kyseisessä verkko-osoitteessa 1–N kertaa. Kun osoitteen sivu on käyty läpi, selaaja siirtyy sivulta linkitettyyn seuraavaan osoitteeseen. [16, s. 18]

Indeksin muodostaminen alkaa yksittäisten sanojen jäsentämisellä ja erottamisella selatuista sivuista ja dokumenteista. Sitä varten hakupalvelulla on käytössä ennalta määritetty sääntö, mitkä sanat indeksiin otetaan mukaan ja mitkä jätetään pois. Vastaavasti tarkastelu täytyy ulottaa myös pienempiin yksiköihin ja merkkeihin, kuten numerot ja symbolit. Sivun URL lasketaan kuuluvaksi sivuun, ja sen sisältämät sanat indeksoidaan myös. [8, s. 79–80]

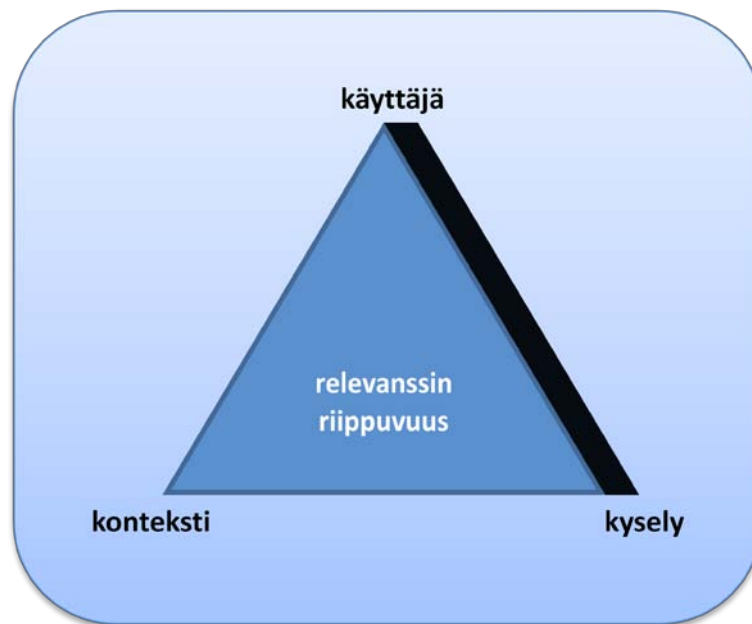
Indeksin sisältämien sanojen tai merkkien rajoittamista hallinnoidaan niin sanotulla stopword-listalla, johon kerätään kielessä hyvin yleisesti toistuvia sanoja. Tällaisia sanoja ovat tyypillisesti pronominit, olla-verbi ja apusanat (partikkelit). Esimerkiksi tässä työssä tarkastelun kohteena olleessa hakupalvelussa on suomen kielen partikkeli ”ja” suodatettu hakutuloksista pois. Tällä pyritään rajoittamaan indeksin kokoa ja siten tehostamaan sen toimintaa ja ylläpidettävyyttä. Vaikka stopwordit jätetään pois indeksistä, niiden sijaintitieto tallennetaan ja ne ovat tarvittaessa palautettavissa kyselytuloksiin käyttämällä tarkennetun haun operandeja. Monissa hakupalveluissa stopword-listoja ei käytetä tai niitä hyödynnetään osittain, koska ne estävät kokonaiseen lauseeseen pohjautuvan kyselyn tekemisen. [8, s. 80]

Hakupalveluiden kehittyneempiin piirteisiin kuuluu niiden kyky huomioida sanojen taivutusmuotoja (word stemming). Tällä tarkoitetaan hakuindeksin ominaisuutta tarjota enemmän tuloksia huomioimalla käytetyn hakusanan taivutusmuodot. Indeksiin tallennetaan käytetystä merkkijonosta sanan perusmuoto ja taivutukset. Tällöin perusmuodon lisäksi tarvitaan tieto siitä, missä muodossa sana on esiintynyt eri dokumenttiyhteyksissä. [8, s. 80] Toiminto perustuu sanan morfologiseen prosessointiin, missä sana palauteaan perusmuotoonsa tai vartaloonsa ja siitä johdetaan sanalle eri taivutusmuodot. Morfologialla tarkoitetaan lingvistiikassa kielen taivutusjärjestelmää [21, s. 43].

Best betillä tarkoitetaan indeksistä erilleen poimittuja hakusana – hakutulos -pareja, joilla voidaan tarjota käyttäjälle suoraan todennäköisintä ratkaisua hänen kyselynsä. Tämä perustuu laajoissa hakupalveluissa tilastolliseen seurantaan, jossa vertaillaan käytettyjä hakusanoja ja niiden perusteella valittuja dokumentteja. Näistä kaikkein suosituimmiksi osoittautuneet yhdistelmät voidaan tallentaa erikseen indeksiin. Kun käyttäjä syöttää best bet -hakusanan hakukoneelle, se osaa tarjota tuloksissa ensisijaisesti hakusanaan liitetyn tuloksen ennen muita tuloksia. Yksittäisten verkkopalveluiden, kuten esimerkiksi yrityksen intranetin hakutoiminnon kehittämisessä best betiä voidaan hyödyntää tehokkaasti ohjaamalla käyttäjät tarkoituksellisesti tietyllä hakusanalla tietyn dokumentin luokse. [22, s. 286]

2.7 Hakutulosten esittämistapa tiedonhaun ohjaajana

Loppukäyttäjä hakee tietoa syöttämällä hakukoneelle yhden tai useampia hakusanoja ja saa takaisin listauksen parhaiten kyselyä vastaavista dokumenteista. Hakutulosten järjestäminen mihinkään absoluuttiseen järjestykseen relevanssin mukaan on käytännössä mahdotonta. Tämä johtuu siitä, että käyttäjät ovat erilaisia tarpeiltaan, koulutukseltaan ja kulttuuriselta taustaltaan. Prosessiin vaikuttaa käyttäjän taustatekijöiden lisäksi haettavan asian kontekstiriippuvuus, eli se mihin laajempaan kokonaisuuteen se liittyy. Vastaavasti se, miten kysely muotoillaan merkkijonoksi, vaikuttaa aina lopputulokseen. Hakukone voi käytännössä hyödyntää ainoastaan käyttäjän sille antamaa merkkijonoa hakutulosten tuottamiseen ja järjestämiseen. [8, s. 79] Kuvassa 7 on esitetty relevanssin arvioinnin riippuvuuteen vaikuttavat taustatekijät.



Kuva 7. Hakutulosten relevanssin arviointiin vaikuttavat riippuvuudet

Hakutulossivu ohjaa käyttäjän valintaa paitsi järjestämällä hakutuloksen listaukseksi, myös esittämällä valintaa helpottavaa informaatiota. Tällaista tietoa voivat olla kohteeseen liittyvät metatiedot tai lyhyt tiivistelmä hakusanan asiayhteydestä dokumentin sisällössä. Koska hakutulosten listaaminen vaatii aina laskentatehoa palvelimelta, vaikuttaa tuloslistauksen optimointi järkevällä algoritmilla myös palvelinkuormaan. Vähemmän kuormitettu palvelin palauttaa hakutulokset nopeammin ja luotettavammin. Hakukone voi hyödyntää myös välimuistia yleisimpien kyselyiden tallentamiseen, mikä vähentää kuormitusta ja nopeuttaa toimintaa. [8, s. 86]

Taulukossa 2 on esiteltyä tärkeimpiä menetelmiä ja tekniikoita, joita hakukoneet voivat hyödyntää relevanssin arviointiin hakutulosten listauksessa.

Taulukko 2. Hakutulosten järjestämisen menetelmiä [8]

Menetelmä	Vaikutus
Sanojen esiintymistiheys, Term Frequency (TF)	Dokumentin sijoitus nousee tuloksissa sen mukaisesti, kuinka usein hakusanat esiintyvät sivulla ja kuinka lähekkäin toisiaan ne ovat.
Lausevastaavuus	Hakukone nostaa sellaiset sivut korkealle, joista löytyy täsmälleen hakua vastaava merkijono.
Linkitykset sivulle	Sivu on sitä tärkeämpi, mitä enemmän siihen on linkityksiä. Lähdesivujen tärkeys on myös pisteitetty ja vaikuttaa kohdesivun arvoon listauksessa.
Linkkitekstin hyödyntäminen	Linkitykseen käytetty teksti lasketaan osaksi kohdesivun sisältöä ja nostaa siten sivun arvoa tuloslistauksessa.
URL-analyysi	Mikäli hakusana esiintyy www-osoitteessa, nostetaan sivu tuloslistauksessa korkeammalle. Esim. Googlessa haku ”rakennustieto” tuottaa 1. tulokseksi www.rakennustieto.fi:n.
HTML-rakenteen painotukset	Sivun sijoitusta tuloksissa voidaan nostaa suunnittelemalla sen rakenne ja typografia hyvin ja hyödyntämällä koodin ominaisuuksia. Avainsanojen sijoittaminen otsikkotyyleiksi määriteltyihin teksteihin sekä tekstikoroistusten käyttö nostavat sijoitusta.
Dokumentin luonti- tai muokkauspäivämäärä	Tuloslistauksissa voidaan painottaa dokumentin sijoitusta sen mukaan, koska sitä on viimeksi muokattu.

Tulosten listaus vaatii hakukoneelta menetelmiä, joilla tuloksen relevanssi voidaan mahdollisimman luotettavasti johtaa dokumentin ominaisuuksista. Koska käyttäjät tutkitusti siirtyvät harvoin 2. hakutulossivulle tai siitä eteenpäin, on korostetun tärkeää pystyä tuottamaan käyttäjän kannalta parhaat tulokset mahdollisimman ylös listauksessa. [8, s. 81–88]

Kohdesivun arvottaminen siihen kohdistettujen linkitysten perusteella on käytännössä aiemmin kerrottua paljon monimutkaisempi ja kehittyneempi menetelmä, mutta sen tarkempi esittely ei kuulu tämän insinööriyön rajaukseen. Käytännössä dokumentin sijoittuminen tuloslistauksessa on aina näiden menetelmien yhdistelmä, eikä mitään ehdotonta ja varmasti toimivaa ratkaisua ole. Hyvä dokumentin rakenteen suunnittelu ja listausmenetelmien huomioiminen on kuitenkin avainasemassa hakukonenäkyvyydessä. Menetelmien hyödyntämistä kutsutaan *hakukoneoptimoinniksi* (SEO, search engine optimization). [8, s. 82] Lisäksi kaupallisilla hakupalveluiden tarjoajilla on omia mene-

telmiään, joita pyritään suojelemaan paitsi liikesalaisuuksina, myös sen vuoksi, ettei niitä käytettäisi väärin [8, s. 89].

Ideaalissa käyttötilanteessa hakutulosten ensimmäiseksi listattu dokumentti vastaa käyttäjän tarvetta. Koska tilanne ei useinkaan ole niin hyvä, joutuu käyttäjä tekemään valintoja dokumenttien välillä. Valinnan helpottamiseksi käyttäjälle esitetään erilaista informaatiota kohdedokumenteista helpottamaan valintaa. Alkeellisimmassa esitysmuodossa hakutuloksissa ei tarvitsisi listata kuin linkit dokumentteihin. Käytännössä erilaiset hakupalvelut näyttävät dokumentista ainakin seuraavat tiedot:

- **Dokumentin nimi**, joka on yleensä dokumentin metatiedoissa esitetty otsikko (title). Verkkosivulla se on HTML-koodissa otsikkoa varten varatun title-tagin sisällä oleva teksti ja muissa dokumenteissa tiedoston sisäisessä metatiedossa esitetty otsikko.
- **Lyhyt tiivistelmä** hakusanan esiintymisestä dokumentin sisällön tekstissä, missä hakusana on esitetty yleensä korostettuna (highlight).
- **Dokumentin URL**, josta voi päätellä yleensä suhteellisen paljon informaatiolähteestä, esimerkiksi onko kyse verkkokaupasta vai tuotevalmistajasta.

Kuvassa 8 on Googlen hakutulossivu, jossa on tehty kysely hakusanalla ”rakennustieto”.

Google rakennustieto

Haku: kaikkialta internetistä suomenkielisiltä sivuilta sivuja maa

Internet Näytä valinnat... Tulokse

Rakennustieto
 RT-kortistot, LVI-kortistot, kirjoja ym. rakennustietoutta.
www.rakennustieto.fi/ - [Välimuistissa](#) - [Samankaltaisia](#)

Tuotteet ja palvelut	Ajankohtaista
Verkkokauppa	Mediapalvelut
Yhteystiedot	ALE Rakennuskirjakaupoissa ...
Haku	Asiakastuki

[Lisää tuloksia kohteesta rakennustieto.fi »](#)

Rakennustieto
Rakennustiedon ajantasainen tietopalvelu on rakentamisen ja kiinteistönpidon vastuuhenkilöiden vankka tuki sisältäen tarvittavat tiedot suunnittelusta ...
www.rakennustieto.fi/kortistot/ - [Välimuistissa](#)

Rakennustiedon verkkokauppa
 Tervetuloa **Rakennustiedon** verkkokauppaan ... **Rakennustieto Oy** • PL 1004 (Runeberginkatu 5), 00101 Helsinki • Puh. 0207 476 400 (v) • Ly: 0113188-9 • Legal ...
kauppa.rakennustieto.fi/ - [Välimuistissa](#) - [Samankaltaisia](#)

Tarviketieto.Net
 RT Tarviketieto Net -palvelun sisältö perustuu rakennustarvikeyritysten teettämiin RT- ja LVI-tuotekortteihin, jotka laaditaan **Rakennustieto Oy:n** ...
www.tarviketieto.net/ - [Välimuistissa](#) - [Samankaltaisia](#)

Rakennustiedon Lehdet
 Tervetuloa **Rakennustieto** Lehdet -portaaliin ... 2010 **RAKENNUSTIETO** LEHDET - PL 1004 00101 Helsinki - Puh. 0207 476 400 ...
www.rakennustietolehdet.fi/ - [Välimuistissa](#) - [Samankaltaisia](#)

Kuva 8. Googlen hakutulossivu

Sivulla on esitettyä edellä mainittujen perusasioiden lisäksi myös hakutulosten kokonaismäärä sekä mahdollisuus rajata haku koskemaan vain suomenkielisiä sivuja. Lisäksi haku osaa tarjota suoran linkityksen rakennustieto.fi:n alisivuille. Sivun yläosassa on myös linkki ”Näytä valinnat”, joka tarjoaa mahdollisuuden tarkentaa tulosten esittämistä suodattamalla tuloksista pois osia ja rajaamalla tulosten esittäminen tiettyyn alueeseen (engl. scope). Taulukossa 3 on esitetty erilaisia Googlen tarjoamia vaihtoehtoja hakutulosten rajaamiseen ja suodattamiseen.

Taulukko 3. Googlen hakutulosten suodatusvaihtoehdot

Hakutulosten rajaaminen, suodattaminen ja kohdistaminen Googlessa				
Kohdistaminen sisältöön	Ajallinen rajaaminen	Historia	Näkymä	Kuvanäkymä ja kielikäänös
kaikki tulokset videot blogit teokset keskusteluryhmät	milloin tahansa uusimmat tulokset viim. tunti viim. 24 tuntia viim. viikko viim. vuosi oma aikaväli	kaikki tulokset käydyt sivut käymättömät sivut	vakionäkymä wonder wheel	vakiotulokset kuvia sivulta sivujen esikatset käännetty haku

Käytännössä tulosten suodattaminen perustuu indeksiin tallennettuun tietoon, jonka mukaan esitettyjä hakutuloksia voidaan rajata. Taulukossa esitetty ajallinen rajaaminen voidaan toteuttaa kohdedokumenttien muokkauspäivämäärän mukaan. Muokkausajankohta on tallennettuna indeksiin, ja indeksistä voidaan poimia esimerkiksi vain viimeisimmän viikon aikana muuttuneet.

3 Tiedonhakuprosessin laadun kehittämisen menetelmiä

3.1 Tiedonhaun kehittäminen edistää tietämyksenhallintaa

Tiedonhakuprosessi on keskeinen osa yrityksen tietämyksenhallintaa. Sitä kehittämällä yritys tavoittelee samalla parempaa tuottavuutta, lisäarvoa ja kilpailukykyä. Tietämyksenhallinnalla tarkoitetaan yrityksen toiminnan ja kannattavuuden kannalta keskeisen tiedon kokoamista, varastointia, jakamista ja käyttöä. Yrityksen sisäinen tieto on kuitenkin usein heikosti järjestettyä, eikä mitään systemaattista tapaa tiedon esittämiseen rakenteisessa muodossa (structured form) ole käytössä.

Tietämyksenhallinnan näkökulmasta tavallinen sanahaku on usein tehoton, koska se palauttaa liikaa epäolennaista informaatiota eikä osaa yhdistellä tietoa. Tiedon hyödyntämiseen kuluu tarpeettomasti aikaa, ja sen ylläpitäminen ajantasaisena on haasteellista. [23, s. 2–4] Verkkopalvelun hakutoiminto on käyttäjälle luonteva tapa löytää ja käyttää tietoa. Tiedon systemaattinen järjestäminen jollakin suunnitelmallisella menetelmällä antaa mahdollisuuden myös tiedonhaun kehittämiseen.

Sanahakuja on kehitetty nykyisistä lähtökohdista hyödyntämällä keinoälyä ja laskennallista lingvistiikkaa, mutta se on osoittautunut haasteelliseksi. Vaihtoehdokseksi ratkaisuksi on muodostunut verkkosisältöjen esittäminen paremmin koneellisesti luettavassa ja prosessoitavassa muodossa. Luonnollisten kielten merkityksiä tutkiva ja määrittelevä semantiikka (merkitysoppi) on käsitteenä tuotu myös tietojenkäsittelyyn. Sillä pyritään kuvaamaan myös verkkoon tallennetun informaation merkitystä ilman, että käyttäjän tarvitsee tulkita tietoa eri asiayhteyksien kautta. Tavoitteena on koota tiettyyn asiaan liittyvä tieto merkityksen mukaan yhdeksi kokonaisuudeksi. [23, s. 3–4]

Yrityksen tietämyksenhallinnan kannalta tämä merkitsee sisäisten prosessien kuvaamista ja luokittelua. Tuotettavan tiedon pitää olla tallennettuna rakenteisessa muodossa ja kuvattuna siten, että se vastaa mahdollisimman hyvin käyttötilanteiden tarvetta. Tätä varten tiedolle täytyy antaa sellaisia ominaisuuksia, jotka liittävät sisällön johonkin laajempaan kokonaisuuteen ja antavat sitä kautta merkityksen tiedolle. Tiedonhakua tukemaan voidaan luoda sisäinen sanasto ja metatietomalli.

3.2 Semanttinen web tiedonhaun kehittäjänä

Semanttisella webillä (semantic web) tarkoitetaan nykyiseen Internetiin kehitettyä laajennusta, jonka avulla verkon tarjoamalle sisällölle pyritään määrittelemään tarkempi merkitys. Samalla tavoitteena on kehittää ihmisten ja koneiden välistä yhteistoimintaa helpommaksi. [11] Verkkopalveluiden sisältö on perinteisesti suunniteltu pelkästään ihmisten luettavaksi ja käytettäväksi. Sisällön rakenteessa ei ole otettu huomioon selaista mahdollisuutta, missä tietokoneet pystyisivät antamaan merkityksiä toisten tietokoneiden sisältämälle informaatiolle. Niiden rooli rajoittuu esitettävän verkkosivun muodostamiseen ja välittämiseen loppukäyttäjälle, eikä luotettavaa menetelmää sivun sisällön merkityksen kuvaamiseen ole ollut olemassa. [24]

Semantiikan avulla pyritään järjestämään verkkoon tallennettu tieto siten, että se voidaan jäsentää käsiteavaruuksiin tiedon merkityksen mukaan. Tiedon ajantasaisuus ja päivittäminen pyritään varmistamaan automatisoidusti. Perinteisessä verkkopalvelussa käytetty asiasanahaku korvataan laajemmilla kysymys-vastauslauseilla, mikä vastaa inhimillistä tiedonhahmottamista paremmin. Tiedonhakuun voidaan tarvittaessa vastata yhdistämällä tietoa useista eri dokumenteista ja lähteistä uudeksi kokonaisuudeksi. Tietoa voidaan myös kohdentaa rajatulle käyttäjäryhmälle käyttöoikeuksia ja tiedon näkyvyyttä rajaamalla [23, s. 4].

Semanttinen web tarjoaa paremman mahdollisuuden sisällön tehokkaampaan integraatioon kuin nykyinen verkkoteknologia. Multimediasisällöt, kuten teksti, kuva, ääni ja videot, voidaan liittää tehokkaammin yhteen palvelemaan tiedonhaun kokonaisuutta. Koneoppimisen (machine learning) hyödyntäminen tarjoaa mahdollisuuden käyttäjäympäristön personointiin. Hakukoneiden tekniikan kannalta semanttinen web tekee verkkosivustojen ja sisällön selausrobottien yhteistoiminnan tehokkaammaksi. Samalla tiedon relevanssin määrittämiseen tarjoutuu uusia tehokkaampia mahdollisuuksia sosiaalisten verkkojen menetelmiä hyödyntämällä. [8, s. 312] Olennaisinta on, että semantiikkaa hyödyntämällä ei verkossa olevia resursseja kytketä toisiinsa ainoastaan yhdellä tavalla käyttämällä hyperlinkkejä, vaan siihen on useita erilaisia tapoja [11].

Semanttisella haulla tarkoitetaan semanttisen webin hakusovellusta. Sillä pyritään kehittämään perinteisten hakumenetelmien tuloksia liittämällä hakuprosessiin mukaan semanttisen webin tarjoamaa dataa ja tiedon rakenteellisuutta. Se ei sellaisenaan ole hyödyksi ja sovellu tavanomaiseen hakuun, jossa tavoitteena on löytää jokin tietty dokumentti. Semanttisessa haussa käyttäjä kuvailee lauseella kohdetta, josta hän haluaa saada lisätietoa, ja saa lopputuloksena yksittäisen dokumentin sijaan laajemman kokonaisuuden. Se sisältää kuvauksen haetun tiedon asiayhteydestä ja helpottaa käyttäjää hyödyntämään tietoa. [11]

Lauseen käyttämisessä on ongelmana sanojen monimerkityksellisyys ja sanojen suhteet toisiinsa. Esimerkiksi sanalla ”voi” on suomen kielessä useita merkityksiä. Sanan esiintyminen lauseessa asettaa vastaavasti muut lauseessa esiintyvät sanat aina eri kontekstiin riippuen sen tarkoituksesta. Tämän vuoksi lauseen jokaiselle sanalle täytyy pystyä johtamaan kaikkien merkitysten yhdistelmät siten, että kyselynä esitettyä lausetta voidaan pitää uskottavana ja loogisena. Tässä muodossa esitetty semanttinen haku on soveltuva reaali maailman ilmiöiden, kuten ihmisten, paikkojen ja tapahtumien, hakemiseen. [11]

3.3 Ontologian käyttö sisällön kuvaamisessa

Ontologiat ovat yleisiä määrittelyjä mille tahansa entiteetille, henkilölle tai asialle. Niitä tarvitaan erilaisten termien ja termistöjen välisten suhteiden selvittämiseen. Sen vuoksi niiden merkitys on ratkaiseva koneymmärteisille verkkosivuille, joita semanttisessa webissä käytetään. [25, s. 80] Ontologialla on perinteisesti tarkoitettu filosofian tutkimushaaraa, joka tutkii olemassaoloa ja olemisen käsitettä. Tietojenkäsittelytiede ja keinoälytutkimus ovat valinneet ontologian omaan käsitteistöönsä kuvaamaan termien välisiä suhteita muodollisesti kuvaavaa kokonaisuutta. Kokonaisuus on yleensä kuvattavissa yhtenä tekstidokumenttina. [24]

Ontologia on tiettyä käyttötarkoitusta varten rakennettu malli, joka kuvaa jotain käsittealuetta tai -joukkoa. Se voidaan esittää ”kuvana maailmasta koneille”. Koska kyse on mallista, sitä voidaan pitää käsitejoukon abstraktiona, jolle on olemassa mahdollisia vaihtoehtoja. Tämä on huomioitava ontologian muodostamisessa ja määriteltävä sen

sisältö käyttötarkoituksen mukaan. Tärkeää on myös pyrkiä ennakoimaan sellaiset tulevaisuudessa tarvittavat laajennukset, jotka ovat jo luontihetkellä ennustettavissa. [23, s. 206] Ontologioita luodaan yleensä aihepiireittäin, vaikka pyrkimys luoda yleisen tason ontologia on myös olemassa. Se on kuitenkin haasteellista, koska määrittelyyn osallistuvat ryhmät saattavat päätyä erilaisiin ratkaisuihin keskenään. Tämän vuoksi on tarve kuvata ontologioita toisiinsa suhteutettuina yhdenmukaisilla menetelmillä. Kuvaamista varten on kehitetty erilaisia kuvauskieliä, kuten OWL (Web Ontology Language). [25, s. 80] Se on kuvattu tarkemmin tämän insinööriyön luvussa ”3.6 OWL verkkosisältöjen kuvausmenetelmänä”. Ontologian kehitysprosessi voidaan Antonioun ja van Harmeleinin mukaan jakaa kahdeksaan päävaiheeseen, jotka on kuvattu taulukossa 4.

Taulukko 4. Ontologian muodostamisvaiheet [23, s. 205–209]

Ontologian muodostamisen kahdeksan päävaihetta	
1. <i>Rajaus</i>	Kehitettävälle ontologialle rajataan kohderyhmä, sisältö ja käyttötarkoitus. Ideaa voi testata kysymyksillä, joihin ontologian pitäisi vastata.
2. <i>Uusiokäyttö</i>	Hyödynnetään mahdollisuuksien mukaan jo olemassa olevia ontologioita uuden lähtökohtana.
3. <i>Termien listaus</i>	Kirjataan vapaasti ja jäsentämättömästi kaikki ne relevantit termit, joiden oletetaan liittyvän ontologiaan.
4. <i>Taksonomian määrittely</i>	Listatut termit järjestetään hierarkkisesti luokkiin siten, että ne muodostavat taksonomisen kokonaisuuden. Taksonomialuokkien periytyvyyden täytyy olla aitoa, jotta kuvauskielet (esim. RDFS, OWL) pystyvät hyödyntämään sitä.
5. <i>Ominaisuuksien määrittely</i>	Taksonomialuokkiin liitetään ominaisuudet. Luokkien ominaisuuksien täytyy vastaavasti olla aidosti periytyviä, kuten itse luokatkin.
6. <i>Piirremääritykset (facet)</i>	Ominaisuuksille määritetään kardinaliteetti (määrän pakollisuus), vaadittavuus ja suhteelliset ominaisuudet. Tämä vaihe vaatii OWL-kuvauskielen käyttöä.
7. <i>Ilmentymien määrittely</i>	Luokista ja ominaisuuksista muodostetaan kombinaatiot objektien ilmentymien kuvaamiseksi ja sisällytetään objektit ontologiaan. Koska niiden lukumäärä on tyypillisesti moninkertainen verrattuna luokkiin, tehdään määrittely ja objektien keruu yleensä koneellisesti analysoimalla lähdetietokantaa.
8. <i>Poikkeamien tarkistus</i>	Ontologiasta tarkistetaan epä johdonmukaisuudet ja korjataan ne.

Muodostamisen lopputuloksena syntyy taksonomia, joka on jaettu objektiluokkiin. Taksonomia määrittelee luokkien väliset suhteet. Se on jaettu alaluokkiin, jotka perivät (inheritance) ylempien luokkien ominaisuudet. Esimerkiksi *osoite* voi olla tyypiltään *paikkaan* liittyvä ja *postinumero* voidaan vastaavasti määrittellä ainoastaan paikkaan liittyväksi. [24]

Yleinen suomalainen ontologia YSO on kehitetty korvaamaan aiempi Yleinen suomalainen asiasanasto YSA. Sen tarkoituksena on tehdä kansallisista tietopalveluhankkeiden sisällöistä paremmin koneellisesti ymmärrettäviä ja näin tehostaa tiedonhaun kehittämistä semanttisen webin menetelmin. [26, s. 1–2]

3.4 Metatiedon hyödyntäminen dokumenttien kuvailussa

Metatiedolla tarkoitetaan tietoa kuvailevaa tietoa eli tietoa tiedosta. Sillä voidaan kuvata yksittäistä tietuetta, dokumenttia tai sen osaa, tai laajemmin tietueiden muodostamaa loogista kokonaisuutta. Metatieto antaa ylätasoa merkityksen tiedolle ja tehostaa sen käyttöä ilman, että koko sisältöä täytyy avata loppukäyttäjälle. Tiedon kuvailulla voi olla useita erilaisia käyttötarkoituksia, mutta yhteistä niille kaikille on, että dokumentti pyritään luokittelemaan ja sijoittamaan johonkin kontekstiin suhteessa muihin dokumentteihin tai resursseihin, mikä helpottaa loppukäyttäjää hyödyntämään tietoa. [27, s. 1–2]

Loppukäyttäjä ymmärretään tässä semanttisen webin käsitteistön avulla, ja se voi näin ollen olla ihmisen lisäksi myös kone. Metatieto on verkkoresurssien (URI) kuvailua [18, s. 113]. Resurssi voi olla esimerkiksi *www-sivu*, *palvelu*, *kuva* tai *ladattava tiedosto*. Olennaista on, että resurssilla on tunnus, joka yksilöi sen.

Metatiedon luokitus

Metatieto luokitellaan sen tyypin mukaan kolmeen eri ryhmään. Kuvaileva eli deskriptiivinen metatieto yksilöi loppukäyttäjälle resurssin ja kuvaa sen sisältöä. Kuvaileva rooli on keskeinen juuri tiedon haussa ja hakutulosten esittämisessä, koska se kertoo selkeimmin tiedonhakijalle dokumentin tarkoituksen. Sisällön kuvaus ja kohdentaminen metatiedon avulla on myös olennainen tekijä hakuindeksin kehittämisessä ja tulos-

ten painottamisessa. Kuvailevaa metatietoa on tyypillisesti dokumentin nimi tai otsikko, tekijä, tiivistelmä ja asiasanat. [27, s. 1]

Rakenteellista eli strukturaalista metatietoa käytetään resurssin koostamisessa ja yhdistelyssä. Se kertoo, mistä elementeistä tai objekteista resurssi (esim. dokumentti) koostuu. Rakenteellista metatietoa voidaan käyttää esimerkiksi osoittamaan, miten dokumentin sivut yhdistetään kappaleiksi. Hallinnollinen eli administratiivinen metatieto on resurssin tarvitseman ja sisältämän teknisen informaation ylläpitoon käytettyä kuvailevaa tietoa. Siihen kuuluvat esimerkiksi tiedostotyyppi, dokumentin luontiaika ja -tapa, tiedostokoko sekä käyttöoikeuksista kertova informaatio. [27, s. 1]

Erillisinä metatietotyyppinä käytetään usein käyttöoikeuksien hallintaan ja resurssin säilyttämiseen liittyvää metatietoa. Säilyttämiseen liittyvän metatiedon tarkoitus on kuvailla dokumentti niin, että sen arkistointi ja pitkäaikainen säilytys on mahdollista. [27, s. 1]

Metatiedon tehtävät

Tärkeä syy kuvailevan metatiedon käyttöön ja tuottamiseen on relevantin tiedon löytäminen. Se tarjoaa mahdollisuuden järkevien kriteerien käyttöön resurssien etsimisessä ja yksilöinnissä. Näin voidaan yhdistää toistensa kaltaiset tai toisiinsa liittyvät resurssit. Vastaavasti sitä voidaan käyttää erottamaan toisistaan poikkeavat resurssit ja antamaan tiedolle tarkka, yksilöity sijainti. [27, s. 2]

Verkkopalvelujen käyttö on luonut tarpeen luokitella sivustoja niiden sisältöjen mukaan ja koostaa portaaleja tiedonhaku varten. Tällä tarkoitetaan sähköisten aineistojen ja resurssien järjestämistä tiedonhaun tarpeiden mukaan. Aiemmin vastaavia palveluita on tuotettu manuaalisesti perinteisten verkkosivujen ylläpidon menetelmillä. Metatietoa voidaan hyödyntää tuottamalla sivustot dynaamisesti tietokantoihin varastoidun metatiedon perusteella. [27, s. 2]

Metatieto mahdollistaa tiedon vaihtamisen ja siirron teknologia-alustaltaan erilaisten järjestelmien välillä ja sitä kautta niiden keskinäisen yhteiskäytön. Kuvattu resurssi on näin myös koneellisesti tulkittavissa, eikä ainoastaan ihmisen ja koneen välisessä vuorovaikutuksessa. Järjestelmien keskinäinen ja saumaton yhteiskäyttö on mahdollista metatietoskeemojen, yhtenevien tiedonsiirtoprotokollien ja skeemojen älykkään linkityksen avulla. Yhteiskäytössä metatietoa hyödynnetään erityisesti järjestelmien välisessä tiedonhaussa. [27, s. 2]

Metatietoa hyödynnetään myös verkkoresurssin sijainnin kuvaamisessa ja yksilöinnissä. Skeemaan kuuluu elementti, johon on sisällytetty kuvaus objektin tai resurssin paikasta. Tyypillisesti verkkoresursseja on kuvattu käyttämällä suoraa URL-osoitetta, mikä on aina absoluuttinen viittaus kohteeseen. Mikäli kohteen sijainti muuttuu, ei URL enää toimi. Käyttämällä metatietoa resurssin sijainnin kuvaukseen voidaan uusi sijainti välittää eri palveluille, jotka hyödyntävä kyseistä resurssia. [27, s. 2]

Tulevaisuuden ongelma on tiedon säilyttäminen luku- ja käyttökelpoisena. Digitaalisen tiedon perusongelmana on sen sidonnaisuus käytettyyn teknologiaan. Aiemmin käytössä ollut tiedostomuoto ja sen käyttöön vaadittavat ohjelmistot ja laitteistot eivät välttämättä toimi enää tietyn ajan kuluttua. Lisäksi tiedostot saattavat vahingoittua. Metatietoa voidaan hyödyntää tiedon säilyttämiseen ja palauttamiseen tulevaisuudessa. Objektin tai resurssin kuvailussa käytettyyn skeemaan voidaan liittää arkistointia ja säilyttämistä kuvaavia elementtejä. Niihin voidaan tallentaa tietoa kohteen fyysisistä ominaisuuksista ja sen käytöstä ja toiminnasta, jotta tiedon palauttaminen ja sen käytön emulointi olisi mahdollista tulevaisuudessa. [18, s. 124–125]

Metatietoa hyödynnetään verkkopalvelun hakutoiminnossa hakujen tarkkuuden parantamiseen. Dokumentin kuvauksesta voidaan valita ne tiedot, jotka parhaiten edesauttavat sen löytymistä tiedonhaussa. Nämä tiedot toimivat *hakuelementteinä*. Ne ovat teki-jöinä hakutulosten relevanssin kehittämisessä ja arvioinnissa. Loppukäyttäjälle tämä merkitsee nopeutunutta tiedonhakuprosessia ja sitä kautta ajan säästöä. [18, s. 118]

Dokumenttien validoinnissa voidaan käyttää hyväksi metatietoa liittämällä dokumentin laatukriteerit käytettyyn metatietoon. Jos dokumentin sisältämä tai siihen liitetty metatieto ei vastaa järjestelmän vaatimuksia, ei dokumenttia voi tai saa liittää järjestelmään. Yrityksen sisäisessä dokumentinhallinnassa tätä voidaan hyödyntää esimerkiksi pakottamalla loppukäyttäjä täyttämään riittävä määrä metatietoja dokumenttiin, ennen kuin sen saa tallentaa intranetiin. Pakollisiksi merkityt metatiedot voidaan määrittää esimerkiksi sisäisen tiedonhaun kriteerien perusteella. [18, s. 115]

Metatiedon rakenne ja muodostaminen

Metatietomalli perustuu skeemaan, joka koostuu elementeistä ja niiden ryhmistä. Kul-lakin ryhmällä on rajattu käyttötarkoitus, joka voi olla esimerkiksi tietyyntyyppisen re-surssin kuvailu. Skeeman semantiikka määräytyy elementtien määrittelyn ja tarkoituk-sen kautta. Elementeille annetaan arvot, jotka antavat sisällön mallille. Lisäksi siihen voidaan liittää tarkentavaa tietoa siitä, missä muodossa sisältö pitää esittää ja mitkä ovat sallittuja arvoja. Kuvauksessa on yleensä tarkennus siitä, mitä syntaksia sen tuottami-seen ja lukemiseen tarvitaan (esim. XML, RDF, SGML). Metatietoskeema voi olla myös riippumaton syntaksista. [27, s. 2]

Metatiedon tuottamisprosessi vaihtelee käytettävän järjestelmän, kuvattavan sisällön luonteen ja sitä tuottavan organisaation mukaan. Hallinnollista metatietoa tuotetaan pal-jon automaattisesti tai tuottajana voi olla esimerkiksi aineiston digitalisoinnista vastaava henkilöstö. Kuvailevan metatiedon tuotanto on yleensä sen kohteena olevan resurssin luoja tai ylläpitäjän tehtävä. Lisäksi voidaan prosessissa hyödyntää informaatioita tai vastaavia ammattilaisia, jotka ovat riittävän perehtyneitä kohteen sisältöön. [27, s. 10] Tekijöiden motivointi omien julkaisujensa metatietojen tuottamiseen on osoittautunut kuitenkin haasteelliseksi. Vaikka erilaisia helpottavia työkaluja on kehitetty kuvailua varten, on tekijöitä ollut vaikeaa saada omaksumaan metatiedon hyödyntämisen ajatus. [18, s. 119]

Tunnetuimpia vakioituja metatietomalleja on Dublin Core. Se on alun perin kehitetty kirjoittajille, jotta he voisivat kuvailla julkaisujaan verkkopalveluihin paremmin sovel-tuvalla tavalla. Kirjastopalveluiden käyttämät tietomallit eivät soveltuneet verkkokäyt-

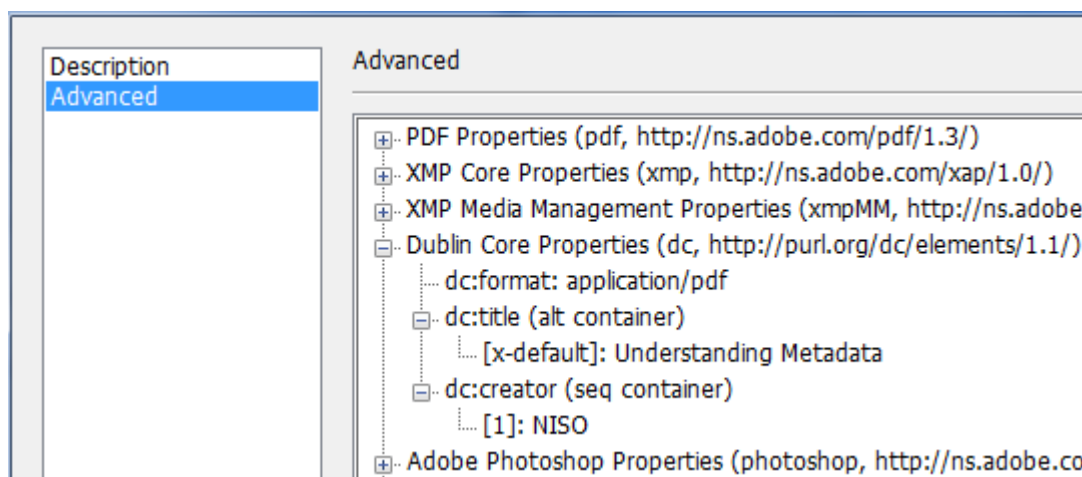
töön, joten oli tarve kehittää yksinkertaisempi ratkaisu tehostamaan julkaisujen verkko-käyttöä. Dublin Core sisältää 15 erilaista elementtiä julkaisun sisällön kuvaamiseen. Taulukossa 5 on esimerkki mallilla kuvatusta kirjasta.

Taulukko 5. Dublin Core -metatieto-esimerkki [lähteen 27, s. 3 mukaan]

Julkaisun kuvaaminen Dublin Coren avulla	
Title =	”Rakennusperinnön tulevaisuus - Poimintoja teemavuoden aiheista”
Creator =	”Nikula, Riitta”
Creator =	”Tuominen, Laura”
Creator =	”Nevanlinna, Tuomas”
Creator =	”Uuskallio, Irma”
Subject =	” rakennettu ympäristö”
Description =	”Rakennusperinnön tila yhteiskunnassamme ja siihen liittyvät haasteet sekä rakennusperinnön tulevaisuus”
Publisher =	” Ympäristöministeriö”
Publisher =	” Rakennustieto Oy”
Date =	” 2004-03”
Type =	” Text”
Format =	” print”
Identifier =	”http://kauppa.rakennustieto.fi/fi/productcard.asp?productid=JUL5119”
Language =	” fi”

Vaikka Dublin Core sisältää 15 elementtiä, eivät ne esimerkin mukaisesti kaikki ole pakollisia. Lisäksi tietyt elementit voivat saada useita arvoja, kuten esimerkiksi tekijä (Creator). [27, s. 3]

Useat ohjelmat ja tiedostotyypit tarjoavat suoran tuen metatiedon hyödyntämiseen. Kuvassa 9 on Adoben pdf-formaatin metatietotaulu, jossa on avattuna tiedoston sisältämä Dublin Core -informaatio.



Kuva 9. Pdf-tiedoston Dublin Core -metatiedot

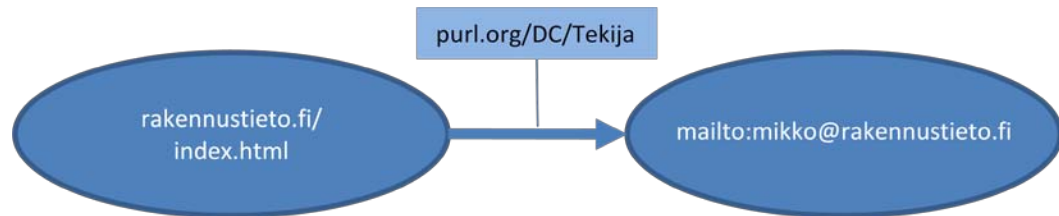
Kuvasta on nähtävissä viittaus Dublin Coren määrittelevään verkkoresurssiin, kohteen formaatti, otsikko ja tekijä. Metatietoja voi tarvittaessa muokata samalla sovelluksella.

3.5 Semanttisen webin sisällön mallintaminen

Resource Description Framework eli RDF on metatietomalli ja kuvausmenetelmä, jolla verkossa julkaistuihin dokumentteihin eli *verkkoresursseihin* (web resources) voidaan liittää informaatiota dokumentin sisällön merkityksestä. Se hyödyntää XML:a (eXtensible Markup Language) sisäisenä alemman tason syntaksina omien osiensa prosessointiin. [8, s. 312] Dokumentilla tarkoitetaan tässä paitsi yksittäistä www-sivua, myös muita sellaisia dokumenttityyppejä, jotka tukevat jotain tapaa sisällön esittämiseen rakenteisesti mallintamalla. Tällaisia ovat muun muassa pdf-tiedosto ja useat toimisto-ohjelmien tiedostotyypit. Kehittyneillä metatietomalleilla, kuten Dublin Core, on myös suoraan hyvä RDF-yhteensopivuus [28]. RDF on W3C-konsortion julkaisema suositus [29].

Dokumenttien sisällön kuvaus perustuu informaation jäsentämiseen tietomallilla (graph data model), jossa muodostetaan subjekti–predikaatti–objekti-kolmikoita (RDF-triples). [29] Mallin mukaisesti jollakin *resurssilla* (subjekti) on *ominaisuus* (predikaatti), joka saa *arvon* (objekti). RDF:n tietomallissa subjekti vastaa tarkasteltavaa yläkäsitettä tai niiden joukkoa. Esimerkiksi lause ”Auto on väriltään punainen” voidaan jakaa osiin siten, että auto on resurssi, jolla on ominaisuutena väri. Ominaisuudella taas on vastavasti arvo, joka on punainen. Predikaatti kuvaa subjektin ja objektin välistä suhdetta.

Kuvassa 10 on havainnollistettu RDF-kolmikon osien suhdetta esimerkillä, joka voidaan lukea ”mikko on verkkoresurssissa rakennustieto.fi sijaitsevan index.html:n tekijä”.



Kuva 10: RDF-kolmikon rakenne [30, s. 2]

Ominaisuudelle voi olla annettu useampia kuin yksi arvo. Resurssien yksilöimiseen käytetään URI:a (Uniform Resource Identifier), jolla osoitetaan tietyn resurssin sijainti verkossa. Verkkosoite, kuten rakennustieto.fi, on URI:n erikoistapaus. Muodostettuja kokonaisuuksia kutsutaan sanastoiksi (RDF vocabulary). Niihin kootaan keskitetysti tiettyyn kontekstiin liittyvää metatietoa, joka ohjaa tiedonhakua. Sanaston kuvaamiseen käytettyä kieltä kutsutaan RDF-skeemaksi, joka on kuvauskielen semanttinen laajennus ja lyhennetään yleisimmin RDFS. [30, s. 1–2, 5]

RDF on semanttisten verkkojen kehittämisessä keskeinen menetelmä. Sen hyötyjä ovat sisällön erottaminen esitystavasta ja mahdollisuus käsitellä koneellisesti metatietoa siten, että se on ihmiselle ymmärrettävää luonnollisen kielen kautta. Sitä kautta se tarjoaa kehittyneempiä mahdollisuuksia tiedonhakuprosessin kehittämiseen kuin nykyiset, keskenään heikosti kommunikoivat verkkopalvelut ja hakukoneet. [8, s. 312]

3.6 OWL verkkosisältöjen kuvailukielenä

OWL (Web Ontology Language) on kuvauskieli, jota käytetään ontologioiden muodostamiseen ja määrittelyyn [16, s. 95]. Kuvaamisessa käytettyjen kielten yhteisenä tavoitteena on esittää tarkkoja formaaleja malleja jonkin rajatun aihepiirin käsitteistöstä. Keskeisiä vaatimuksia kielen toimivuudelle ovat

- hyvin muodostettu lauseoppi (syntaksi)
- muodollisesti määritellyt merkitykset

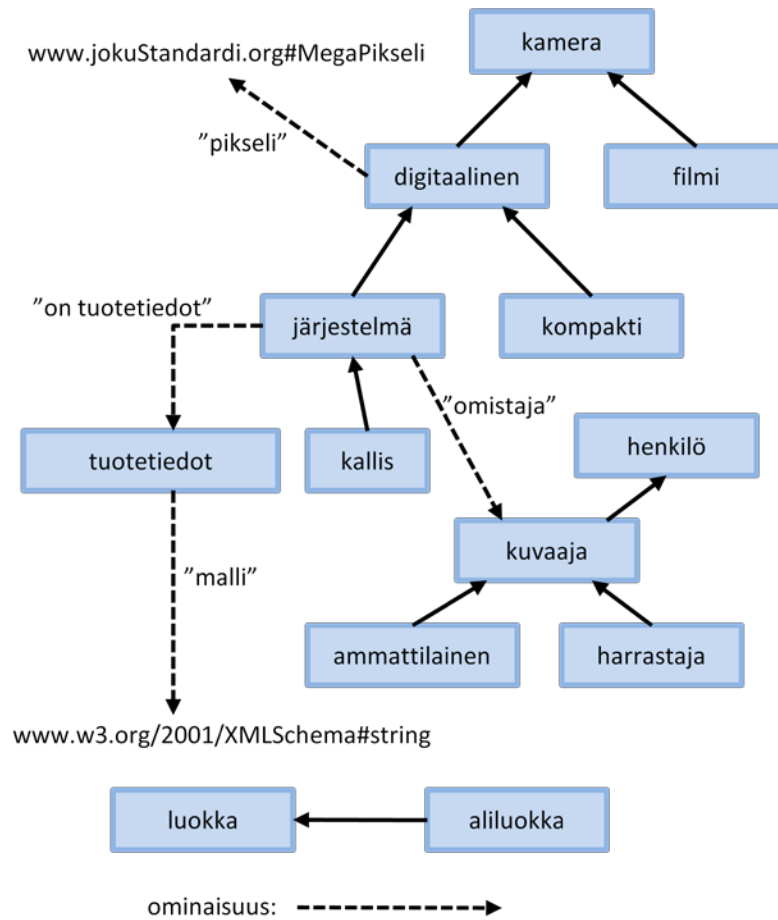
- tehokas tuki päättelylle
- riittävä ilmaisuvoima
- soveltuvuus ilmaisuun.

Lauseopin puhtauden vaatimus on samanlainen kuin ohjelmointikielissä. Se mahdollistaa informaation koneellisen käsittelyn. Tarkasti määriteltyjen merkitysten eli formaalin semantiikan vaatimus kohdistuu tiedon merkityksen tarkkaan kuvaamiseen ja on verrattavissa matematiikan logiikkaan. Semantiikan hyödyntäminen on edellytys vaatimuksena olevalle päättelylle. Kuvailukielen tarjoama tuki päättelyyn mahdollistaa ontologian ja sen sisältämän tiedon johdonmukaisuuden tarkistamisen. Siten voidaan ehkäistä tahattomat virheelliset suhteet muodostettavien luokkien välillä ja luokitella objektien esiintymiä automaattisesti [23, s. 110–111].

OWL on W3C:n suositus, ja siitä on marraskuussa 2009 julkaistu 2-versio, joka lisää uusia ominaisuuksia kieleen ja säilyttää yhteensopivuuden aiemman kanssa [31]. OWL käyttää perustanaan RDF-skeemaa ja lisää siihen käsitteiden ilmaisutapoja. Tämän vuoksi kaikki skeeman luokat ja ominaisuudet ovat käytettävissä myös OWL:ssa. Sen käyttötarkoitus on vastaavasti määritellä luokkia, niiden ominaisuuksia ja keskinäisiä suhteita. Kuvailukielen käytön etu verrattuna pelkkään skeemaan on sen kyky esittää paljon monimutkaisempia ja -puolisempia suhteita objektien välillä. Lopputuloksena voidaan rakentaa sellaisia työkaluja tai web-agentteja, joilla on laajennettu päättelykyky. RDF-skeeman rajoituksien vuoksi OWL:a pidetään ensisijaisensa kuvausmenetelmänä ontologioiden kehittämisessä [16, s. 95].

OWL on jaettu kolmeen eri alikieleen erilaisten lähestymistapojen tarpeen vuoksi. Eri versioita kutsutaan nimillä Full, DL (description logic) ja Lite. Käytännössä kielien erot ovat käytettävien luokkien muodostimien (constructor) hyödyntämisessä ja RDF-yhteensopivuudessa. Full-versio tarjoaa eniten ominaisuuksia ja Lite on vastaavasti helppoin ottaa käyttöön. [23, s. 113–114]

Tiivistettynä esimerkkinä OWL-mallinnetusta ontologiasta on kuvassa 11 esitetty kameran liitettyä käsitteistöä.



Kuva 11. Esimerkki kameran ontologiasta OWL-mallinnuksessa [16, s. 98]

Mallissa on esitetty luokkaan "Kamera" liittyviä käsitteitä. Kaaviosta voidaan poimia esiin sellainen ilmentymä, jossa tarkastellaan Kamera-luokassa esiintyvää digitaalista järjestelmäkameraa. Aliluokasta on erotettu ammattikäyttöön tarkoitettut kalliit kamerat erikseen. Kameralla on ominaisuuksia: sillä on tuotetiedot (esim. kameran malli) ja sen omistaa kuvaaja, joka on joko ammattilainen tai harrastaja. Vastaavasti, koska kamera on tyypiltään digitaalinen, on sillä ylemmältä tasolta periytyvänä ominaisuutena "pikseli". [16, s. 98]

4 Hakutoiminnon kehittäminen käyttäjätestauksen avulla

4.1 Tutkimuksen testimetodit ja arviointikriteerit

Intranetin hakutoiminnon testaamista tarkastellaan tämän insinööriyön yhteydessä laajemmassa kontekstissa verkkopalvelun käyttäjätestauksen yhtenä erikoistapauksena, jolla on selkeämmin rajatut lähtökohdat ja tavoiteltava lopputulos. Työn kohteena ollut hakutoiminto oli osa tietojärjestelmää, ja sen vuoksi tuloksien arvioinnissa ei voi käyttää yksin verkkopalvelun toimivuutta kokonaisuutena arvioivia menetelmiä. Testauksen tuloksien arvioinnissa ei keskitytty pelkästään verkkopalvelun käytettävyyden mittareiden tarjoamiin arviointikriteereihin (esim. Nielsen, Krug mukaan). Ne painottavat tyyppillisesti käyttöliittymäsuunnittelua ja navigointihierarkiaa. Näitä kriteereitä voidaan tässä tutkimuksessa hyödyntää hakusyötteen käyttöliittymän ja hakutulosten esittämisen käytettävyyden arviointiin.

Testauksella pyrittiin selvittämään hakutoiminnon käytettävyydessä ja toimivuudessa esiintyviä puutteita. Lähtökohtana oli kehitystesti, jonka tarkoitus oli löytää loppukäyttäjän kannalta mahdollisimman hyvin toimiva ratkaisu tiedonhaun tehostamiseen intranetissä. Luotettavasti toteutetun käyttäjätestauksen tavoitteiden mukaisesti käytettävyys pyritään aina testauttamaan lopullisilla käyttäjillä todellisessa käyttötilanteessa tai todellisuutta mahdollisimman hyvin jäljittelevässä ympäristössä. Sillä ei tarkoiteta järjestelmän toimituksen loppukatselmusta, jossa voidaan verrata toimintoja määrittelydokumenttiin kirjattuihin vaatimuksiin. [32, s. 276–277]

Hakutoiminnon testaamiseen on olemassa omia erityisiä menetelmiään. Ne käyttävät tuloksen relevanssin erilaisia arviointimenetelmiä järjestelmän toimivuuden mittauksessa. Lähtökohtana arvioinnissa on, löytyykö haluttu dokumentti ja miten korkealle se on sijoittunut tuloslistauksessa. Relevanssin arviointimenetelmiä kuvataan tarkemmin tämän työraportin luvussa 2.5. Koska tarkastelun kohteena ollut järjestelmä oli jokapäiväisessä tuotantokäytössä oleva intranet, ei käytännössä ollut mahdollista rakentaa sellaista testausympäristöä, jossa relevanssia olisi pystytty luotettavasti ja tarkasti mittaamaan. Tämä johtuu pääasiassa siitä, että intranetin sisältö elää koko ajan ja relevantin sisällön ja hakutulosten saannin suhde muuttuu. Lisäksi kaupallisen hakukoneen kaikkia

toimintaan liittyviä piirteitä ei ole julkaistu dokumentoituina, mikä omalta osaltaan haittasi tarkkojen arviointien tekemistä.

Käyttäjätesti valittiin hakutoiminnon kehittämisen perusteeksi, koska järjestelmää käyttävillä työntekijöillä on erilainen tiedonhakarve. Koska käytössä oli jo tuotantokäytössä oleva järjestelmä, oli kyse myös aidossa loppukäyttäjäympäristössä järjestettävästä testistä. Testaustietoa haluttiin kerätä tilastollisesti mahdollisimman hyvin valitulla otoksella, jotta eri näkökulmat ja lähestymistavat tulisi kartoitettua riittävän selkeästi. Testiasettelussa haettavan tiedon relevanssi otettiin huomioon ja pyrittiin huomioimaan myös se, että relevantti tieto on erilaista eri käyttäjäryhmillä. Lisäksi käyttäjätestistä saatiin tietoa yleiseen käytettävyyteen liittyvistä asioista, kuten ulkoasun miellyttävyydestä ja navigoinnin helppoudesta.

Testauksen tukena käytettiin heuristista analyysia ja arviointia siltä osin, kuin se soveltuu hakutoiminnon testaamiseen. Arvioinnin tukena oli järjestelmän käyttöönotossa saatu käyttäjäpalaute, jossa intranetin hakutoimintoa oli käytetty tiedonhakuun. Saadun käyttäjäpalautteen pohjalta laadittiin testauksen tueksi tarkistuslista hakutoiminnon heuristista arviointia varten.

4.2 Testimalli

Testi toteutettiin kaksivaiheisena siten, että ensin arvioitiin olemassa olevan järjestelmän toimintaa käyttäjätestillä ja -kyselyllä. Ensimmäisessä vaiheessa haastateltiin lisäksi edustava otos loppukäyttäjistä toiveiden ja tarpeiden kartoittamiseksi. Testausta varten laadittiin kyselylomake, jossa oli ohjeistuksen lisäksi viisi eri hakutehtävää ja vapaa kommentointiosuus. Tehtävät oli suunniteltu siten, että ne kohdistivat haut intranetin kannalta keskeisiin osa-alueisiin (henkilöstö, sisäiset palvelut, myynti ja markkinointi, tuotanto). Testaajalle ei ollut osoitettuna yhtä niin sanottua ”oikeaa vastausta”, vaan kysymyksenasettelu määritteli tiedonhauulle tarpeen ja intranetistä löytyi varmasti vastaus hakuun. Viimeinen tehtävä oli vapaamuotoinen haku, jossa käyttäjä sai itse määritellä haettavan dokumentin. Tällä pyrittiin kartoittamaan käyttäjien tarvetta erityyppisen tiedon ja dokumenttien hakuun. Vapaa hakutehtävä arvioitiin samoin kriteerein kuin

muutkin tehtävät. Tehtävät olivat rakenteeltaan identtisiä, ja niillä kerättiin tietoa seuraavin perustein:

- käytetyt hakusanat
- relevantin hakutuloksen sijoittuminen listauksessa
- lopputulokseen vaikuttanut valintatekijä
- käyttäjän tekemät turhat valinnat
- dokumentin löytyminen
- käytetty aika
- yleisarvio tehtävän onnistumisesta.

Testitehtävien suunnittelussa lähtökohtana oli niiden vastaavuus todellisten työelämässä kohdattavien tiedonhakutilanteiden kanssa. Suunnittelussa otettiin huomioon toistettavuus, jotta mitattaville tekijöille pystytään tuottamaan riittävä referenssi myöhemmin. [33, s. 185] Kysely lähetettiin sähköisenä lomakkeena kohderyhmän jäsenille. Vastaajat valittiin organisaatorakenteen mukaisesti siten, että edustettuina olivat sisäiset palvelut, tietopalvelu ja asiakaspalvelu painotettuna henkilöstömäärällä. Otoksesta jätettiin pois operatiivinen johto ja IT-osasto, koska niiden mukaan ottaminen olisi saattanut vääristää lopputulosta johtuen liian läheisestä suhteesta tarkasteltavaan järjestelmään. Ennen kyselyn lähettämistä lomake testautettiin pilottikäyttäjällä. Saadun palautteen mukaan lomakkeen sisältöä muokattiin tehtävienannoissa käytettyjen sananmuotojen osalta. Kyselylomake on tämän työn liitteenä (liite 1).

Testin kulku suunniteltiin siten, että käyttäjä

1. avaa intranetin (hakuikkuna on kaikilla sivuilla)
2. lukee tehtävänannon
3. valitsee hakusanan ja syöttää sen hakukoneelle sekä kirjaa lomakkeeseen
4. käy läpi hakutulokset
5. valitsee relevantin tuloksen ja kirjaa sen sijoituksen
6. muokkaa hakusanaa tarvittaessa, jos haluttua tulosta ei tule, ja toistaa vaiheita 3–5, kunnes haluttu tulos löytyy

7. merkitsee tiedon lomakkeeseen, jos tulosta ei kohtuullisessa ajassa löydy
8. kirjaa lopuksi aika- ja yleisarviot suorituksesta
9. kirjaa lomakkeen loppuun vapaan kommentin
10. lähettää lomakkeen XML-datan palautusosoitteeseen ”Lähetä”-painikkeella.

Testitulokset kerättiin automaattisesti Microsoft Excel -taulukkolaskentaohjelmalla jatkokäsittelyä ja laskentatoimenpiteitä varten. Kysely lähetettiin kaikkiaan 14 henkilölle, joista määräpäivään mennessä vastasi 12. Vastausprosentti oli näin ollen 86 %. Yrityksessä työskenteli testausajankohtana 87 henkilöä, joten vastausprosentti edusti 14 %:a yrityksen henkilöstöstä ja otos vastaavasti 16 %:a.

4.3 Tulosten arviointi

Testitulosten arvioinnin tärkein kriteeri oli loppukäyttäjän tarvetta vastaavan hakutuloksen löytyminen siten, että se esiintyy mahdollisimman korkealla hakutulosten listauksessa. Lisäksi tiedonhaun tehokkuutta kuvaa myös se, että käyttäjän on mahdollisimman helppo valita oikea dokumentti tulostuloksesta. Tutkimuksessa käytettyjä muita kvantitatiivisia arviointimenetelmiä olivat mediaaniluokkien hyödyntäminen hakuajan arvioinnissa ja kokonaisarviossa. Näillä tekijöillä pyritään arvioimaan hakutuloksen relevanssia käyttäjän tavoitteleman tiedon suhteen. Relevanssin arvioinnissa painopiste on käyttäjärelevanssissa aihelevanssin sijaan. Aiherelevanssia voidaan periaatteessa mitata vertaamalla esimerkiksi metatietomallin yhdenmukaisuutta vastaavien hakutuloksien frekvenssiin. Tämä ei kuitenkaan kuvaisi loppukäyttäjän tarvetta tarkasti, eikä se siten sovellu pelkästään arvioinnin perustaksi. Käyttäjärelevanssiin pohjautuva arviointi sitoo tutkimuksen selkeämmin käyttäjätutkimuksen arviointimenetelmien perusteisiin [34].

Testitulosten arvioinnin perusteet ja kriteerit voidaan jaotella esimerkiksi seuraavasti (tärkein ensin):

1. hakutuloksen käyttäjärelevanssi (miten hyvin osuma vastaa käyttäjän tarvetta)
2. hakutuloksen aihelevanssi (miten hyvin osuma vastaa käytettyä hakuterminä esim. metatietomallin tai hakuun liitetyn ontologian tai muun sanaston mukaan)

3. hakutuloksien esittäminen (miten käyttäjä pystyy tulosten listauksesta arvioimaan tehdyn haun relevanssin)
4. suorituksen nopeus (miten nopeasti käyttäjä pääsee tavoitteeseensa)
5. yleinen käyttäjäytyväisyys (miten loppukäyttäjä kokee hakutapahtuman toiminnallisuuden kokonaisuutena).

Koska tutkimus pohjautui käyttäjien tekemiin tehtäviin, kohta 2 ei painottunut testituloksissa. Kehittämisvaiheessa havaittiin myöhemmin, että puutteelliset metatiedot ja esitystapa häiritsivät aihe relevanssin mittausta. Tehtävän onnistumisen arvioinnissa käytetyt luokat (1 = epäonnistunut, 5 = onnistunut) toimivat vastineena Nielsenin ongelmien vakavuuden listausasteikolle ja niitä käytettiin paitsi kokonaisuuden, myös yksittäisten tapausten arviointiin [33, s. 103].

Tulosten kvantitatiivisen käsittelyn ja arvioinnin lisäksi hyödynnettiin käyttäjäpalautetta hakutoiminnon kehittämistarpeen arvioinnissa. Kehitysmahdollisuudet oli sidottu käytettävään järjestelmätoteutukseen, eikä ulkopuolista ohjelmointi- tai konsulttiapua ollut käytettävissä. Arviointia täydennettiin soveltuvien osin heuristisesti. Testimalli mahdollisti tapausten uusimisen tehtyjen muutosten jälkeen tietyissä rajoissa ja siten muutosten vaikutusten arvioinnin. Koska kyseessä oli tuotantokäytössä oleva järjestelmä, jonka sisältö elää ja muuttuu kaiken aikaa, ei uusintatarkastuksen toteuttaminen identtisenä muutosten jälkeen ollut enää mahdollista. Tämä johtui pääasiassa siitä, että hakuun indeksoitujen dokumenttien esiintymät moninkertaistuivat kehitysajanjaksona. Tulosarvioinnissa tämä on pyritty ottamaan huomioon painotuksina.

4.4 Heuristiset menetelmät hakutoiminnon kehittämisessä

Kehitystyön menetelmänä hyödynnettiin myös heuristista arviointia täydentämässä käyttäjätestauksen antamia tuloksia. Sillä tarkoitetaan asiantuntijamenetelmää, jossa ei käytetä loppukäyttäjiä testihenkilöinä, vaan arviointi perustuu valistuneen henkilön tekemiin päättelyihin järjestelmän toiminnasta. Heuristisessa arvioinnissa pyritään käyttämään tarkistuslistoja järjestelmän mahdollisten puutteiden löytämiseen. [35, s. 2–3]

Intranetin hakutoiminnon testauksessa heuristiikkaa hyödynnettiin

- sanallisen testipalautteen arvioinnissa
- hakukoneen tiedossa olevien ominaisuuksien systemaattisessa läpikäymisessä
- käytetyn metatiedon ja hakutuloksen vastaavuuden relevanssin arvioinnissa (aiherelevanssi)
- hakuoperandien toiminnan arvioinnissa
- sanojen taivutusmuotojen (stemming) toimivuuden arvioinnissa
- haun yleisen teknisen toimivuuden arvioinnissa.

Heuristiikan hyödyntämisen referenssinä käytettiin tunnettujen hakupalvelujen (esim. Google) toimintaa ja Nielsenin yleistä heuristiikkalista siltä osin, kuin se soveltuu hakutoiminnon arviointiin. Listan kohdista haun testaamiseen soveltuivat erityisesti käyttäjän kielen ja järjestelmän yhteensovittaminen sekä käyttäjälle esitettävän epärelevantin informaation välttäminen käyttöliittymässä. [36]

5 Hakukonfiguraatio ja metatietomalli

5.1 Kehityksen kohteena ollut intranet

Työn kohteena oli Rakennustieto Oy:n intranetjärjestelmä. Se toteutettiin vuosina 2008–2009 ja otettiin käyttöön huhtikuussa 2009. Intranet toteutettiin käyttämällä Microsoft Office Sharepoint Server 2007 Standard -sovellusta (myöhemmin MOSS). Intranetin käyttökieli on suomi, eikä muita valinnaisia tai käyttäjäryhmittäin kohdennettuja kieliä tarvinnut huomioida projektin määrittelyssä. Toteutuksessa oli mukana alihankkija, joka teki määrittelyn pohjalta tarvittavat muutokset ja lisäykset järjestelmään vakioasennuksen mukana tulevien ominaisuuksien lisäksi.

Hakutoiminto toteutettiin järjestelmään kuuluvalla Enterprise Search -komponentilla. Hakutoiminnon tarkempi määrittely ja toteutus eivät olleet osana alihankkijalta tilattua toimitusta, vaan se kytkettiin päälle vakioasetuksilla. Tämän vuoksi oli jo käyttöönottoaiheessa odotettavissa, että haku ei tule sellaisenaan tyydyttämään loppukäyttäjien tarpeita. Hakua varten tarvittiin oma kehitysprojekti, joka tarjosi mahdollisuuden tämän insinööriyön tekemiseen. Tämän työselostuksen terminologiassa on

pyritty käyttämään Microsoftin käyttämiä suomenkielisiä termejä järjestelmästä siinä määrin, kuin se on ollut mahdollista.

Yrityksen aiempi intranet oli teknisesti vanhentunut, eikä sen käytettävyys vastannut enää nykypäivän tarpeita. Dokumenttien hallinta oli hajautettu verkkolevyille, sähköpostijärjestelmän yleisiin kansioihin ja vanhaan intranetiin. Tiedonhaku perustui sovituihin käytäntöihin asiakirjojen sijainnista. Tietomäärän paisumisen myötä se johti hallitsemattomaan tilanteeseen, johon uudella intranetillä haettiin ratkaisua.

Uuden intranetin lähtökohtana oli viestinnän ja dokumentinhallinnan kehittäminen. Uudistuksen keskeiset tavoitteet olivat

- viestinnän tehostaminen
- henkilöstöhallinnon kanssa tapahtuvan asioinnin tehostaminen
- tiedon ajantasaisuuden ja oikeellisuuden varmistaminen
- tiedonhaun tehostaminen
- dokumenttien versionhallinta ja hallittu julkaisu
- yhtenäiset ja yrityskuvaa tukevat dokumenttipohjat
- hiljaisen tiedon keruu ja saattaminen koko organisaation käyttöön
- turhan sähköpostiliikenteen vähentäminen
- sähköpostiliitteiden vähentäminen
- ryhmätyömenetelmien ja -työkulkujen kehittäminen.

Hakutoiminnon kehittäminen toimii luonnollisesti suoraan tiedonhaun tehostajana, mutta sillä pyrittiin edistämään epäsuorasti myös muiden tavoitteiden täyttymistä. Tällaisiksi todettiin ainakin hiljaisen tiedon keruu ja viestinnän tehostaminen.

Tarpeellisen tiedon mahdollisimman nopea löytyminen vapauttaa aina työaikaa varsinaiseen työtehtävään. Verkkopalvelun hakutoimintoa pidetään yleisesti palvelun käytettävyyden kannalta kriittisenä tekijänä.

Projektissa toteutettiin viestintäintranet, johon liitettiin dokumentinhallintaominaisuuksia. Rakennustiedon kehitysvision mukaisesti intranetin kautta pyritään jakamaan sel-

laista informaatiota, josta on koko organisaation henkilöstölle tai ainakin suurimmalle osalle hyötyä jokapäiväisten työrutiinien hoitamisessa sujuvasti ja tehokkaasti. Tiedon keskitetyllä ja avoimella jakamisella pyritään jatkossa koko organisaation tietämyksen kasvattamiseen. Yksittäinen työntekijä voi tällöin koota laajemmasta informaatiovirrasta tarvitsemansa ja hyödyllisenä pitämänsä tiedon. Tavoitteellisena lopputuloksena on motivoituneempi ja innovatiivisempi työntekijä, kuin mitä perinteisellä hierarkkisella viestintämallilla on saavutettu. Hakutoiminnon todettiin olevan tämäntyyppisessä ratkaisussa keskeisessä roolissa, koska oikein toteutettuna se tarjoaa käyttäjälle relevanttia tietoa laajemmassa kontekstissa. Haun avulla loppukäyttäjä voi tavoittaa intranetistä sellaista merkittävää tietoa, mikä jäisi perinteisen mallin mukaan toimittaessa esimerkiksi toisen yksikön sisäiseen viestintään.

Intranetin hyödyntäminen viestintäprosessissa

Tyypillinen tilanne, jossa hyödynnetään intranetiä laajana viestintäkanavana Rakennustiedossa, on uuden kustanteen tai tuotteen ilmestymisilmoitus. Ilmoituksesta laaditaan intranettiin vakiomuotoinen dokumentti. Sen lisäksi laaditaan sisäinen tiedote, jossa on ilmoituksen sisältö referoituna. Ilmestymisilmoitus sisältää julkaisuun tai tuotteeseen liittyvää metatietoa, referaatin sisällöstä, tuote Profiilitiedot ja tuotevastaavan yhteystiedot.

Taulukossa 5 on kuvattu ilmestymisilmoituksen sisällön käyttötavat ja se, miten käyttäjryhmät hyödyntävät tietoa eri työvaiheissa ja -tilanteissa. Viestintäprosessi käynnistyy yleensä sillä hetkellä, kun tuote (esim. kirja) on valmis ja toimitettu varastoon. Tämän jälkeen tuotevastaava laatii ilmestymisilmoituksen, tallentaa sen intranetin dokumenttikirjastoon ja tiedottaa asiasta henkilökunnalle käyttäen intranetin uutispalstaa viestintäkanavana.

Taulukko 5. Viestintäprosessin merkitys eri aikaväleillä ja kohderyhmillä

Ilmestymisilmoitusta hyödyntävät käyttäjäryhmät ja eri ajanjaksot	
Välittömästi tietoa käyttävät	Varasto Tarkistaa logistiikkajärjestelmän tiedot ja aloittaa toimitukset.
	Myynti Saa myyntiluvan ja perustiedot tuotteesta myynnin aloittamista varten.
	Markkinointi Tuotteen julkistamisen käynnistys, markkinointimateriaalin tarkistus, lehdistötiedotteen julkaisu.
Viiveellä tietoa käyttävät	Taloushallinto Toimii tukena tietojen kirjaamisessa järjestelmiin.
	Markkinointi Tiedon hyödyntäminen tuotteen julkistamista myöhemmissä kampanjoissa.
Tietoa pitkällä aikavälillä käyttävät	Kirjatoimitus Uusintapainoksen tai korvaavan tuotteen taustatietona.
	Tuotekehitys Taustatietona aiemmin tiettyä asiaa käsitelleistä tuotteista.
	Asiakaspalvelu Asiakastiedustelut vanhoista tuotteista.

Koska yrityksen työntekijät tarvitsevat ilmestymisilmoituksen tietoja pitkällä aikavälillä, oli tärkeää, että ilmoitukset löytyvät helposti vielä vuosienkin jälkeen.

Periaatteessa koko organisaatio kuuluu välittömästi tietoa hyödyntäviin jollain tasolla, koska koko henkilöstön on hyvä tietää yleisellä tasolla yrityksen tuotevalikoiman muutoksista. Aiemmin ilmestymisilmoitusten informaatio jaettiin sähköpostitse koko henkilöstölle ryhmäpostituksella. Viestissä ei välttämättä ollut omaa sisältöä, vaan asia kerrottiin liitetiedostossa. Tästä oli seurauksena tyypillisiä sähköpostin käytön ongelmia: viesti hävisi informaatiovirtaan, se saatettiin poistaa ensimmäisen lukukerran jälkeen tai siihen ei osattu enää palata myöhemmin, kun tietoa olisi tarvittu. Ilmestymisilmoitus oli löydettävissä tiedostona myöhemminkin sisäverkossa sijaitsevasta tallennuspaikasta, mutta sitä ei osattu hyödyntää riittävän laajasti, eikä siihen voinut kohdistaa hakua.

Tässä viestintäprosessissa saavutettiin intranetiä hyödyntämällä tiedon keskitetyn jakamisen ja tallentamisen ohella sen saaminen hakutoiminnon piiriin. Se palvelee heti uutisen julkaisuhetkestä lähtien eri työrooleissa tuotetta lähestyviä henkilöitä. Tieto on myös navigoimalla löydettävissä sen omasta tallennuspaikasta ja selailemalla uutisar-

tikkeleita taaksepäin. Sanahaku tarjoaa myös nopean ja helpon tavan dokumentin löytämiseen. Tämä ominaisuus korostuu pitkällä aikavälillä, sillä käyttäjä saattaa muistaa vain aihepiiriin liittyviä asiasanoja ja haettavan dokumentin tarkka nimi ja aihe voi olla unohtunut. Lisäksi sanahaku tuottaa käytännössä aina aiheeseen liittyen myös muita dokumentteja, jotka voivat olla relevantteja käyttäjän tiedonhakuprosessin kannalta. Se mahdollistaa käyttäjälle laajan kokonaisnäkömyksen muodostamisen haetusta aiheesta ja toimii tietämyksen lisääjänä. Sen sijaan, että käyttäjä päätyisi alun perin haluamaansa dokumenttiin suoraan navigoimalla, hän voi haun kautta nyt löytää

- uudempaa tietoa asiasta
- taustatietoa aiheeseen liittyen esimerkiksi vanhemmista saman tyyppin tuotteista
- linkittyvää tietoa siitä, mihin kaikkeen aihe liittyy.

Mahdollisuus yhdistää tietoa uudeksi tiedoksi on tietämyksenhallinnan kannalta ratkaisevaa, ja se oli myös tämän työn kohteen kannalta keskeistä.

5.2 Käyttäjätestauksen tulokset

Käyttäjätestauksella tutkittiin intranetin haun käyttöä todellisilla loppukäyttäjillä aidossa käyttöympäristössä. Testauksen tavoitteena oli löytää toiminnon heikkoudet systemaattisesti ja samalla kerätä käyttäjäkokemuksia. Vapaa arviointimahdollisuus toimi lisäksi heuristista tarkastelua ohjaavana tekijänä yhdessä suoran palautteen kanssa.

Tulosten perusteella pystyttiin nopeasti päättelemään, että haku toimii perusasetuksilla käyttöön otettuna sanahakuna välttävästi. Se palautti hakutuloksia, joissa oli relevanssia suhteessa käytettyihin hakutermeihin. Hakutulosten arviointitermein ilmaistuna saanti oli riittävä, mutta tarkkuus ei ollut riittävä. Käyttäjät eivät olleet varauksettomasti tyytyväisiä uuteen toimintoon. Käyttäjäpalautteesta nousi selvästi ongelmallisina esiin

- tulosten otsikoinnin heikkous
- yhdyssanojen virheellinen käsittely
- puutteet sisällön kattavuudessa
- tarve rajata hakualuetta.

Sanallisessa palautteessa todettiin haun toimivan nopeasti, mutta oikean valinnan tekemisen olevan hankalaa ja turhauttavaa. Dokumentin sijaintia kuvaavaa osoitepolkua pidettiin useassa palautteessa luotettavana valintakriteerinä. Tätä käyttäjäkokemusta vahvistaa se, että intranetin sisäinen rakenne toisintaa organisaation toimintoja ja rakennetta. Osoitepolku viittaa silloin myös asiayhteyteen ja osoitteesta ”http://intra/henkilosto/Henkilostoetudet/” voi helposti päätellä sisällön. Mikäli haun kohteena on tiedosto, auttaa osoitepolun lisäksi myös loogisesti nimetty tiedosto oikean hakutuloksen valinnassa. Tämä on kuitenkin selvässä ristiriidassa käyttäjän luonnollisen kielen käytön ja käsitteellisen yksiselitteisyyden kanssa.

Käyttäjätestin kyselyn perusteella saatiin tilastollisin menetelmin arvioitavia tuloksia tehtävien onnistumisesta. Tulokset on esitelty taulukossa 6.

Taulukko 6. Kyselytulosten tilastollinen analyysi

Käyttäjätestin tulokset	
Tehtävän kohta	Tulos
tehtävien suorituksessa löytymättä jääneet dokumentit	18 % (11/60 dokumenttia)
parhaan hakutuloksen sijoitus listauksessa	7
ylimääräisiä valintoja	2,13
keskimääräinen hakuaika (mediaaniluokka)	4 (1–2 min)
arvio kokonaisuudesta (mediaaniluokka, asteikolla 5 = onnistunut, 1 = epäonnistunut)	4

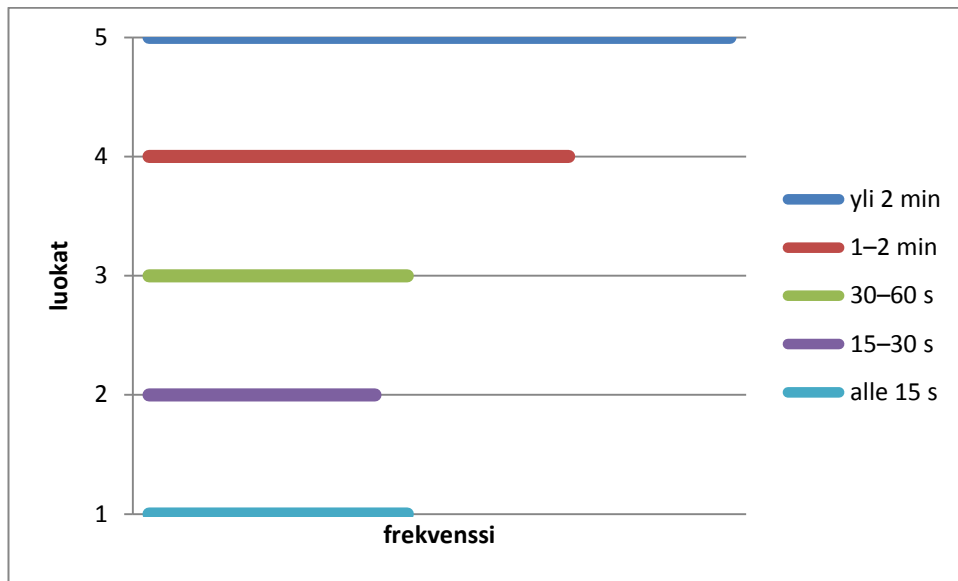
Ensimmäisessä kohdassa on havaittavissa suhteellisen suuri löytymättömien dokumenttien osuus. Tällä tarkoitetaan epäonnistunutta hakutehtävää, jossa käyttäjä ei löytänyt hakukoneen avulla mielestään parhaiten tehtävänasettelua vastaavaa tietoa intranetistä. Tähän ei ole yksiselitteistä ratkaisua, sillä tämä ilmiö ei kohdistunut yksittäiseen tehtävään vaan jakautui tasaisemmin ja painottui pikemminkin suorittajan mukaan. Kaikissa tehtävissä suurin osa testihenkilöistä löysi mielestään relevantin tiedon. Vapaavalintaisessa hakutehtävässä tuli eniten löytymättömiä dokumentteja, koska käyttäjän hakemaa tietoa ei voitu taata löytyväksi intranetistä. Sitä voidaan toisaalta käyttää kuvaamaan tulosten saantia ja sitä kautta sisällön kehitystarvetta. Löytymättömien dokumenttien suuren määrän yhtenä selityksenä voidaan pitää puutteita hakutulosten esittämistavassa, mikä estää käyttäjää valitsemasta relevanttia dokumenttia. Lisäksi käyttäjät eivät jaks

käydä suuria hakutulospäämiä läpi, joten tehtävän keskeyttäminen ennen tuloksen löytymistä voi olla selittävä tekijä.

Hakutulosten listaus ei vastannut käyttäjien odotuksia tiedon relevanssin suhteen, mikä osoitti selvää kehitystarvetta. Myönteistä oli, että keskiarvo osoitti halutun tuloksen esiintyvän ensimmäisellä tulossivulla. Käyttäjätutkimusten mukaan harva käyttäjä viitsii edetä seuraavalle hakutulossivulle siitä huolimatta, että ensimmäinen sivu ei tuota haluttua tulosta [8, s. 25]. Keskiarvon sijoittuminen tulokseen nro 7 on silti nähtävä heikkona lopputuloksena, koska intranet on rajatun käyttäjäryhmän verkkopalvelu ja käyttäjien voi olettaa osaavan hakea tietoa käyttämällä täsmällisempiä termejä kuin esimerkiksi Internet-tiedonhaussa. Tuloksissa oli lisäksi hajontaa tehtävien ja käyttäjien kesken. Se painottui ääripäihin siten, että osa käyttäjistä löysi haluamansa dokumentin, mutta vasta usean hakutulossivun selailun jälkeen.

Tiedonhaussa tulee tyypillisesti turhia valintoja, kun käyttäjä tulkitsee hakutulostuloksen antaman informaation väärin ja luulee löytävänsä haluamansa dokumentin tai tiedon kohdelinkin kautta. Tämä ongelma on kohdistettavissa tulostuloksissa esitetyn metatiedon ja muun oheistiedon laatuun. Listauksessa on hyvä esittää mahdollisimman hyvin dokumentin sisältöä ja tarkoitusta kuvaavaa tietoa. Sen lisäksi tuloksissa pitäisi ilmetä asiayhteys jollain tasolla. Tutkimustuloksen mukaan käyttäjä joutui keskimäärin avaamaan kaksi muuta dokumenttia, ennen kuin löysi haluamansa. Koska käyttäjä joutuu aina arvioimaan, sisältääkö avattu dokumentti tarvittavan tiedon, kuluu sivujen ja tiedostojen aukomiseen tarpeettomasti aikaa. Tämä on selvästi nähtävissä myös tutkimustuloksissa tehtävien ajankäyttöä mittaavassa osiossa.

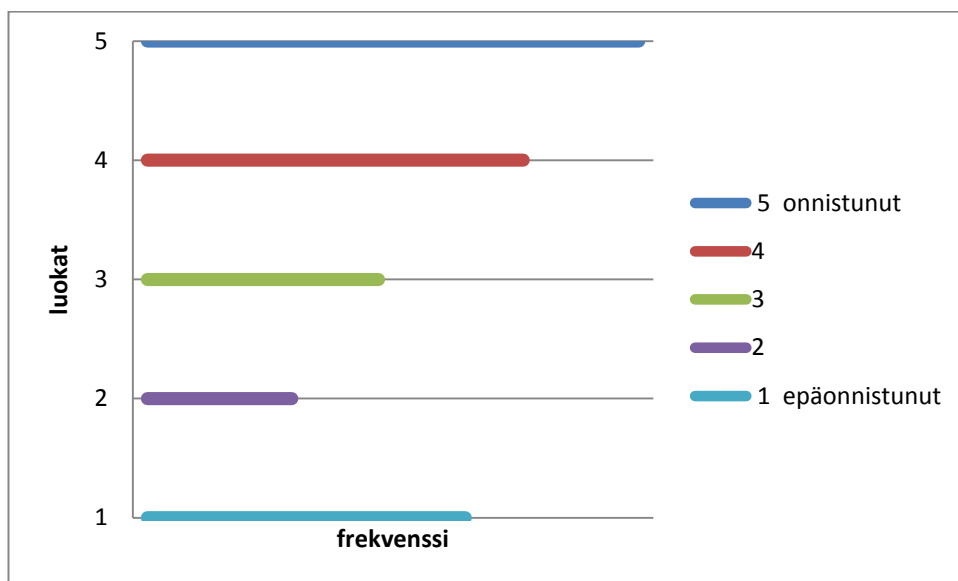
Tehtävien suoritukseen käytetty keskimääräinen aika osoitti selvää kehitystarvetta haun toiminnassa. Kuvassa 12 ovat hakutehtäviin käytetyt arviointiluokat ja niiden suhteelliset frekvenssit kaaviona.



Kuva 12. Hakutehtäviin käytetty aika frekvenssiluokittain

Mediaaniluokaksi muodostui 4 (1–2 min), mistä voi päätellä haun toiminnassa ja käytettyyydessä olevan ongelmia. Tässäkin kohtaa korostuu se, että vaikka verkkopalvelun aihepiiri ja terminologia on käyttäjille jokapäiväistä, se ei korreloi tiedonhaun tehokkuuden kanssa.

Jokaisen tehtävän lopuksi testihenkilöitä pyydettiin antamaan kokonaisarvio tehtävän onnistumisesta. Kuvassa 13 on yhteenveto tyytyväisyysarvioista kaaviona.



Kuva 13. Hakutoiminnon kokonaistyytyväisyys

Kokonaistyytyväisyys nousi yllättävän suureksi mediaaniluokan ollessa 4. Tulos ei sellaisenaan tue hakutehtävien mittaustuloksia, joiden perusteella haun toimintaa ei voinut pitää parhaana mahdollisena. Arvioinnit painottuivatkin ääripäissä sijaitseviin luokkiin. Epäonnistuneiksi tapahtumiksi koetun luokka 1:n frekvenssi on kolmanneksi korkein. Tätä tukee myös sanallinen käyttäjäpalaute, jossa haun ongelmat painottuivat. Suhteellisen korkeaa kokonaistyytyväisyyttä voidaan selittää sillä, että yrityksessä ei aiemmin ole ollut sisäverkkoon kohdistettavaa hakumahdollisuutta. Käyttäjät kokevat jo hakumahdollisuuden selvänä parannuksena vanhaan, ja siihen ei osata vielä välttämättä kohdistaa kriittisiä vaatimuksia.

Tulosten arvioinnin perusteella voitiin nähdä selvä kehitystarve haun käytön nopeudessa ja kokonaistyytyväisyyden nostamisessa. Tehtäviin kulunut aika, turhat valinnat ja löytymättä jääneet dokumentit muodostivat ongelman ytimen. Siihen liittyi myös relevantin hakutuloksen suhteellisen alhainen keskimääräinen sijoittuminen hakutuloksissa. Kehittämistarve kohdistui erityisesti hakutulosten relevanssin nostamiseen paitsi saantia, myös tarkkuutta nostamalla. Lisäksi hakutulosten esittäminen vaati parantamista, koska sillä voidaan vähentää erityisesti turhien valintojen tekemistä ja sitä kautta nopeuttaa tiedonhakuprosessia. Haun toiminnan parantamiseksi päädyttiin

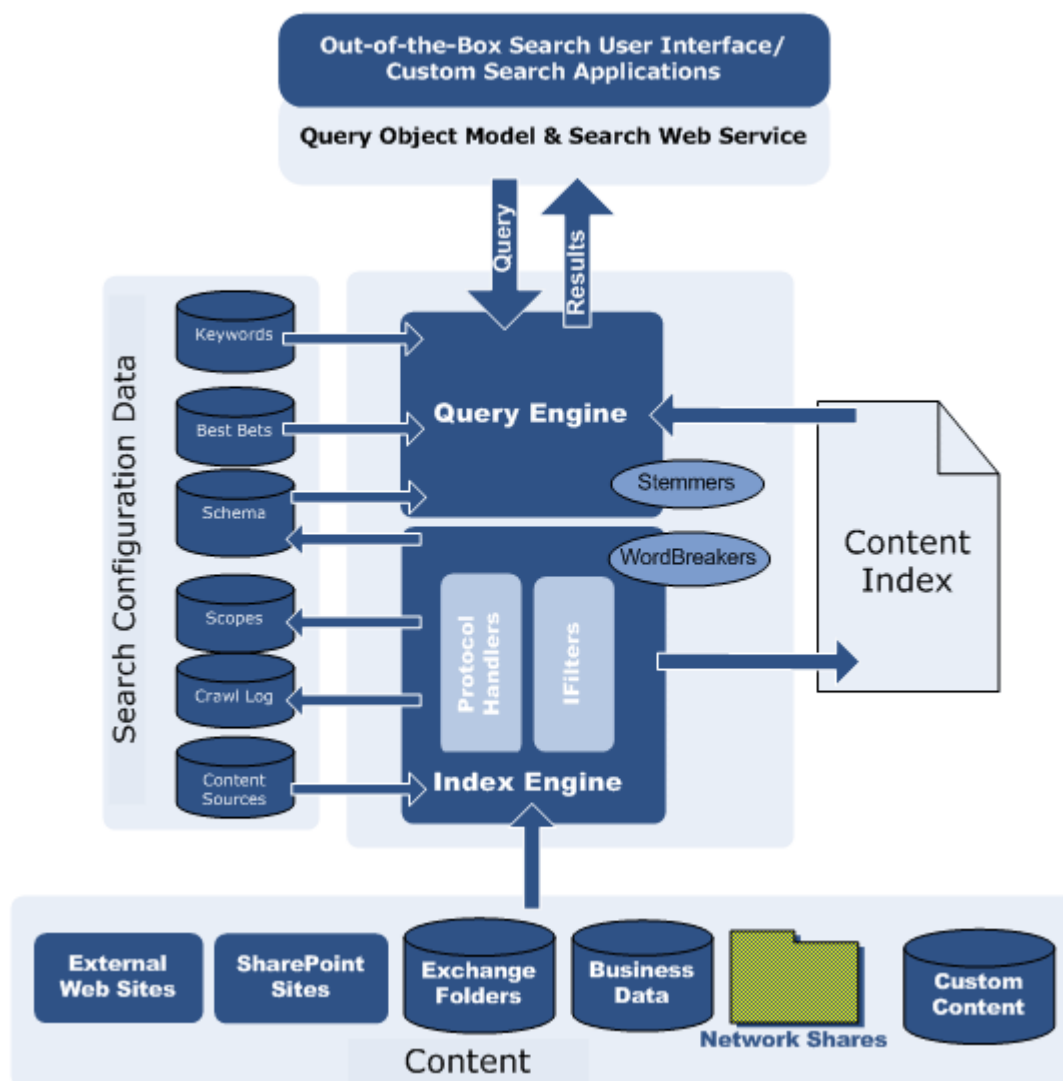
- kehittämään metatiedon käyttöä yrityksessä
- laajentamaan hakuindeksin sisältöä
- kohdistamaan hakuja
- parantamaan haun teknistä toimivuutta.

Näiden parannusten toteuttamiseksi tehdyt toimenpiteet ja muutostarpeiden määrittelyt on kuvattu tarkemmin luvussa ”5.4 Tehdyt toimenpiteet”.

5.3 Käytetyn hakupalvelun sisäinen arkkitehtuuri

Rakennustiedon intranet toteutettiin Microsoftin Office Sharepoint Server 2007:llä, jonka osana on Enterprise Search -komponentti. Se on jaettu palvelu ja toimii toteutetun verkkopalvelun sisällön indeksoijana ja hakukoneena. Kuvassa 14 on esitetty hakutoi-

minnon sisäinen arkkitehtuuri (Microsoftin mukaan). Tässä luvussa käsitellään haun arkkitehtuuria siltä osin, kuin se tämän työn toteutuksen kannalta on olennaista.



Kuva 14. Microsoftin Enterprise Search -arkkitehtuuri [Kuva Microsoft]

Kuvassa ylimpänä on käyttöliittymätaso (User Interface), joka oli tässä ratkaisussa kolmannen osapuolen toteutus. Perushakuna käytettiin sivun ylätunnisteosassa sijaitsevaa Small Search boxia, joka toteutettiin teknisesti samalla tavalla kuin vakioasennuksessa. Hakuun vakiona käytetty Search Center oli piilotettu käyttöliittymästä. Se tuli näkyviin hakutulosten esittämissivuna ja tarkennetun haun käytön yhteydessä. Käyttöliittymän kautta suoritetaan kyselyt ja saadaan hakutulosten palautus hakukoneelta (Query Engine).

Haun ytimenä toimii kaksi keskeistä osaa: hakukone ja indeksoija (Index Engine). Hakukone vastaa kyselyjen käsittelyn logiikasta ja relevanssin mukaisesta tulosten järjestämisestä. Indeksoija selaa sille kohteeksi määritellyn sisällön ja tallentaa sen kuvan oikeassa reunassa esitettyyn sisältöindeksiin (Content Index). Indeksoijan keskeiset osat ovat Protocol Handler ja iFilter. Protocol Handler mahdollistaa tietyssä sijainnissa sijaitsevan sisällön löytämisen, eli käytännössä sitä tarvitaan esimerkiksi URL:n ymmärtämiseen. iFilter taas osaa indeksoida kyseisessä sijainnissa olevan resurssin sisällön. Hakukone tuottaa myös sanojen taipumusmuotojen hyödyntämisen hakutuloksissa (Stemmers), ja vastaavasti indeksoija osaa jakaa sanat osiinsa (Word Breakers).

Kuvan vasemmassa laidassa on koottuna hakukonfiguraation muokkaamiseen käytettävissä olevat menetelmät. Tässä työssä hyödynnettiin menetelmiä seuraavasti:

- **Keywords** Avainsanalistan muokkaaminen suosittujen hakujen pohjalta
- **Best Bets** Suositeltavien hakutulosten kytkeminen avainsanoihin
- **Schema** Metatiedolla kohdennettu haku
- **Scopes** Haun rajaus määriteltyyn indeksin osaan
- **Content sources** Perussivuston ulkopuolisten sisältölähteiden liittäminen indeksiin.

Kuvan alalaidassa on esitettyinä mahdollisia sisältölähteitä. Tämän työn osalta indeksin sisältölähteinä hyödynnettiin intranetin sivustoja ja sisäverkossa sijaitsevia levyjakoja.

5.4 Tehdyt toimenpiteet

Testauksen vapaassa kommentointiosuudessa yleisimmäksi käyttäjien mainitsemaksi yksittäiseksi ongelmaksi osoittautui hakutulosten otsikointi. Testauksessa havaittiin, että hakukone palautti tuloksina runsaasti epärelevanttia materiaalia ennen haluttua tulosta, vaikka hakutermin oli huolellisesti valittu. Hakutuloksen otsikko toimii linkkinä itse kohteeseen, ja se tulee kohdedokumentin metatiedon otsikkokentästä (title). Koska yrityksessä ei ollut käytössä dokumentinhallintaa ja sen myötä systemaattista tapaa liittää metatietoa dokumentteihin, oli otsikkoina osittain täysin sattumanvaraisia tekstejä tai do-

kumentin sisältöä muuten huonosti kuvaavia otsikoita. Kuvassa 15 on esimerkki hakutuloksesta, jossa dokumentti on saanut otsikokseen numeron 11.



Public Folders\allPublic Folders\Ilmestymisilm.\ 14.5.2007 Rakennustieto on kustantanut Tuote: Tekijä: Julkaisijat: Muoto (paperi, CD, Internet): Tuoteprofilointi (Juha): Ammattilaisille 1 2 3 1 ...

http://intra/ajankohtaista/kaikkilomakkeet/Documents/ilmestymisilmoitus_yleinen.doc - 56KB - Kristiina Lehtimäki - 19.3.2009

Kuva 15. Huonosti otsikoitu dokumentti hakutuloksessa

Otsikkona ollut numero ei kuvannut sisältöä millään tavalla. Selvitystyön jälkeen kävi ilmi, että alkuperäisessä dokumentissa oli ensimmäisenä kohtana päivämääräkenttä, josta numero 11 oli joskus dokumentin ensimmäisen käyttökerran yhteydessä tallentunut otsikoksi. Esimerkiksi Microsoft Word ja monet muut ohjelmat lisäävät automaattisesti otsikkokenttään tietyn merkkijonon, tässä tapauksessa se oli dokumentin alun tekstiä. Tässä tapauksessa samaa dokumenttia oli käytetty pohjana pitkän aikaa. Kun vanhoja dokumentteja tuotiin intranetin dokumentinhallintaan sisään, oli haun kannalta tuloksena tietyillä hakusanoilla linkkejä pelkästään eri dokumentteihin nimeltä ”11”. Tilanne korjattiin hyödyntämällä dokumenttikirjastojen metatieto-ominaisuuksia ja lisäämällä dokumentteihin kuvaavat otsikot. Kuvassa 16 on esiteltynä täsmälleen sama dokumentti, johon on korjattu otsikko. Sen jälkeen haku on testattu uudelleen hakusanalla ”ilmestymisilmoituspohja”, ja tulos on järkevä.



Public Folders\allPublic Folders\Ilmestymisilm.\ 14.5.2007 Rakennustieto on kustantanut Tuote: Tekijä: Julkaisijat: Muoto (paperi, CD, Internet): Tuoteprofilointi (Juha): Ammattilaisille 1 2 3 1 ...

http://intra/ajankohtaista/kaikkilomakkeet/Documents/ilmestymisilmoitus_yleinen.doc - 56KB -

Kuva 16. Korjattu otsikointi hakutuloksessa

Tuloksen relevanssi on noussut siten, että käytetyllä hakusanalla se oli sijalla 2 ja lyhyemmällä muodolla ”ilmestymisilmoitus” se sijoittui sijalle 4. Aiemmin sijoitus jäi pois 10 parhaan listalta, koska hakusanan esiintyminen otsikossa on voimakkaasti painotettu


tuloslistauksessa ja otsikkona ollut numero 11 ei tässä tapauksessa liittynyt mitenkään haettavaan tietoon.



Toinen käyttäjätestauksessa esiin noussut ongelma oli relevantin sisällön puuttuminen hakutuloksista. Tämä tuli esiin käyttäjäpalautteessa. Käyttäjät pyrkivät hakemaan tietoa, jonka olemassaolosta he olivat varmoja ja tiesivät, missä dokumentissa sen pitäisi esiintyä. Tällaista tietoa oli muun muassa laatujärjestelmän sisällössä ja pdf-muodossa intranettiin tallennetuissa tiedostoissa. Yrityksen laatujärjestelmä sijaitsi intranetin ulkopuolella, eikä se sen vuoksi sisältynyt perusasennuksen hakuindeksiin. Niin ikään pdf-tiedostot eivät kuuluneet vakiona hakukoneen indeksoijaan määriteltyihin tiedostotyyppiin.


Intranetin hakuindeksiä muokattiin siten, että siihen lisättiin uutena sisältölähteenä yrityksen laatujärjestelmä. Se sijaitsi sisäverkossa olevalla tiedostopalvelimella, ja indeksiin konfiguroitiin sisältölähteen sijainti ja selaussäännöt. Indeksien päivittämisen jälkeen tarkistettiin pistokokeilla ratkaisun toimivuus. Toinen hakuindeksilajennus oli pdf-tiedostojen liittäminen selattavaksi sisällöksi. Sitä varten täytyi asentaa iFilter-komponentti, joka mahdollisti pdf-tiedostojen sisällön näkymisen indeksoijalle. Yrityksessä on paljon pdf:inä tallennettua tietoa, koska se on kustannus- ja media-alalla yleinen tiedostotyyppi. Asennusta varten tarvittiin Adoben verkkosivuilta erillinen asennusohjelma, joka asensi tarvittavan DLL-tiedoston indeksoijan käytettäväksi. Lisäksi pdf piti määritellä rekisteriasetuksissa sallituksi tiedostomuodoksi ja siihen liitettiin kuvake, joka toimii hakutuloksissa ja tiedostolistauksissa tiedostotyyppin indikaattorina.


Vaikeimmin korjattavaksi ongelmaksi osoittautui hakukoneen tapa käsitellä suomen kielen yhdyssanoja. Tämä tuli esiin käyttäjätestauksessa, kun hakutuloksissa esiintyi haettuun tietoon nähden runsaasti täysin epärelevantteja tuloksia. Ongelmaa ei pystytty rajaamaan käyttäjäpalautteen perusteella, ja sitä selvitettiin heuristisella arvioinnilla käymällä läpi hakukoneen eri vaihtoehtoja käsitellä hakusanaa. Syyksi todettiin hakukoneen väärää tapa käsitellä suomen kielen yhdyssanoja. Kuvassa 17 on esitetty hakusanalla ”työaika” tehty kysely. Tuloslistauksessa esitetystä kuvailevassa tiedossa ei


esiinny käytettyyn hakusanaan nähden lainkaan relevanssia. Hakutuloksia on palautunut kuitenkin 768, mikä on erittäin paljon ottaen huomioon indeksin laajuus.


työaika  Tarkennettu haku

Tulokset **1-10** / noin 768. Haku kesti 0,25 sekuntia. Tulosten lajittelu: Asiayhteys | Näyttöperuste: Muokkauspäivämäärä |  Haluan ilmoituksen |  RSS

 **Ritsa**
Intranet Tervetuloa, Palaute Sivukartta Laatujärjestelmä Tarkennettu haku Sivuston toiminnot Henkilöstö > Ritsa ...
<http://intra/henkilosto/ritsa> - mosshallinta - 4.10.2009

 **Rakennustieto**
Rakennustieto TYÖOHJE versio: 3.0 1 Laatujärjestelmä 10.7.2001 laatija: THE, RO TUS22-1
TUS 22-1 LEHDEN SISÄLLÖN SUUNNITTELU Vain ...
<file:///rati8/yhteiset/laatujarjestelma/tus221> lehden sisällön suunnittelu.doc - 22KB - Merja Hakkarainen - 22.11.2001

 **RAKENNUSTIETOSÄÄTIÖ RTS /RAKENNUSTIETO OY**
RAKENNUSTIETOSÄÄTIÖ RTS /RAKENNUSTIETO OY sisäinen lausuntokierros SISÄINEN LAUSUNTOKIERROS
Lausunto aika _____ pvm Lausuntopyynnön _____ lähettäjä ...
<file:///rati8/yhteiset/laatujarjestelma/tit55-l2> sisäisen lausuntokierrosen lomake.doc - 21KB - Raija Lahtinen - 28.4.2004

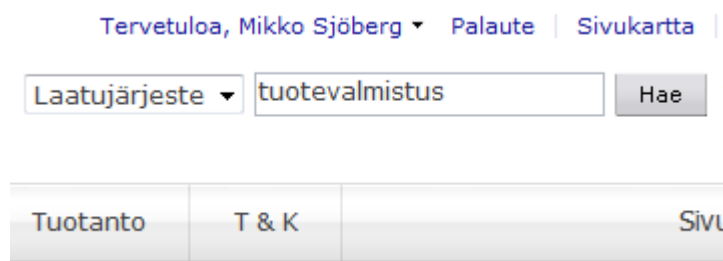
 **TIT55**
Rakennustieto TEHTÄVÄOHJE versio 3.0 1 Laatujärjestelmä 16.10.2001 laatija: MLA, SLS, LV TIT55
TIT55 LAUSUNTOMENETTELY, ...
<file:///rati8/yhteiset/laatujarjestelma/tit55> lausuntomenettely, sisäinen.doc - 32KB - Merja Hakkarainen - 5.10.2009

Kuva 17. Yhdyssanahaun tuottama epärelevantti hakutulos

Yhdyssanojen käyttäytymistä erikseen testaamalla todettiin, että hakukoneen sanavartaloiden palauttamiseen käytetty word breaker jakaa yhdyssanan osiinsa ja tekee haun myös erillissanojen yhdistelmällä, joka tässä tapauksessa on ”työ” JA ”aika”. Tämä tuottaa ymmärrettävästi paljon enemmän hakutuloksia, mutta alkuperäisen kyselyn kannalta ne eivät ole relevantteja. Tämä ongelma voitiin kiertää käyttämällä lainausmerkkejä hakusanan ympärillä, mikä pakotti hakukoneelle lähetettävän merkkijonon käsiteltäväksi juuri siinä järjestyksessä, kuin se on kirjoitettu. Lisäksi havaittiin, että jostain syystä haku toimi oikein, kun se toistettiin tarkennetulla haulla täsmälleen samassa muodossa ja käyttämättä mitään tarkennetun haun parametreja. Ongelman selvittämiseksi tutkittiin järjestelmän kieliasetukset ja asennettiin suomen kielen ominaisuudet hakuun uudestaan. Se ei kuitenkaan auttanut, ja selvittäminen jatkui dokumentaation,

verkkoyhteisöjen ja Microsoftin tuen avulla. Ratkaisua ei kuitenkaan löytynyt tämän työn aikana, ja asia selvitetään myöhemmin. Käyttäjiä pyritään ohjeistamaan, miten välttää ongelma jatkossa. Tutkitusti hakua tarkentavien lisäominaisuuksien käyttö ei kuitenkaan ole suosittua, joten käyttäjien ohjeistamista ei voi pitää ongelman ratkaisuna. Pikaisena korjauksena kokeiltiin JavaScriptillä toteutettua ratkaisua, jossa lainausmerkit pakotettiin hakumerkkijonon ympärille, mutta siitä aiheutui enemmän ongelmia muissa hakusanayhdistelmissä ja se jätettiin ottamatta käyttöön.

Testipalautteessa esiintyi toive mahdollisuudesta kohdistaa haku rajattuun osaan intranetiä. Tätä varten kokeiltiin haun kohdistamista pelkästään laatujärjestelmään ja tiettyä metatietoa sisältävään sisältöön. Laatujärjestelmään suunnattu rajausta toteutettiin hakukoneen käyttämällä sisältöalueen kohdistamisella. Käyttöliittymässä olevaa hakujen rajaamiseen käytettyä alasvetovalikkoa muokattiin siten, että laatujärjestelmä tuli uudeksi vaihtoehdoksi ja järjestelmän sisällä se liitettiin indeksiin määriteltyyn vastaavaan sisältörajaukseen [kuva 18].

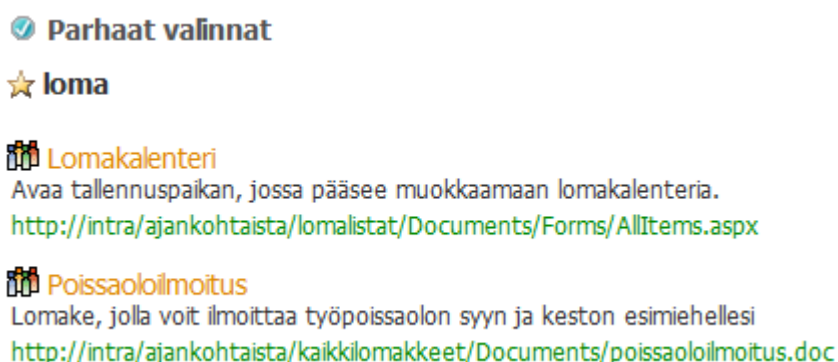


Kuva 18. Laatujärjestelmään kohdistettu hakurajaus

Metatietoon perustuvaa kohdistusta tutkittiin mahdollistamalla haku dokumentin ominaisuuden mukaan. Testitapaukseksi valittiin dokumentin kategoria, joka kuvaa, mihin laajempaan organisaation sisäiseen luokkaan dokumentti kuuluu. Kohdistus toteutettiin järjestelmään kuuluvalla ”Metatietojen ominaisuuksien yhdistämisellä”, jolla tarkoitetaan dokumenttiin liitetyn metatietokentän ja hakuindeksissä vastaavasti indeksoidun metatiedon kytkemistä. Sen seurauksena käyttäjä voi liittää hakuun tarkentavan määrään, esimerkiksi ”kategoria:Sisäiset palvelut”, joka hakee käytetyllä hakusanalla ainoastaan kategoriaan Sisäiset palvelut kuuluvista dokumenteista. Koska metatiedon hyödyntäminen oli tämän insinööriyön aikana vasta alkamassa tilaajayrityksessä, ei tätä

ominaisuutta päästy tutkimaan ja hyödyntämään laajasti. Se testattiin rajoitetusti intranetin työtilaosiossa, johon metatietomalli saatiin työn suorituksen loppuvaiheessa liitettyä. Hakua ohjaavana tekijänä metatiedon hyödyntäminen siirrettiin siten jatkokehitykseen, kun malli saadaan kehitettyä ja otettua laajasti käyttöön.

Intranetin hakutuloksia oli mahdollista kohdentaa tiettyyn dokumenttiin muokkaamalla avainsanalistaa, joka on indeksiin erikseen tallennettu hakusanalista esimerkiksi yleisimmin käytetyistä tai muuten tärkeiksi arvioiduista hakusanoista. Niitä varten voitiin määrittää parhaiten vastaavat dokumentit, jotka tulevat hakutuloksissa pakotetusti ensimmäisiksi ennen yleisesti relevanssin mukaan järjestettyjä dokumentteja. Nämä ovat niin sanottuja Best Bet -tuloksia, joita tämän työn puitteissa päädyttiin käyttämään suosituimpien hakusanojen tilaston perusteella. Kuvassa 19 on esitetty sanalla ”loma” tehty haku, jonka tuloksiin tulee listaukseen ensin ”Parhaat valinnat” -otsikon alle pakotetut tulokset ennen varsinaista tuloslistausta.



Kuva 19. Hakutulosten pakottaminen Best Bet -menetelmällä

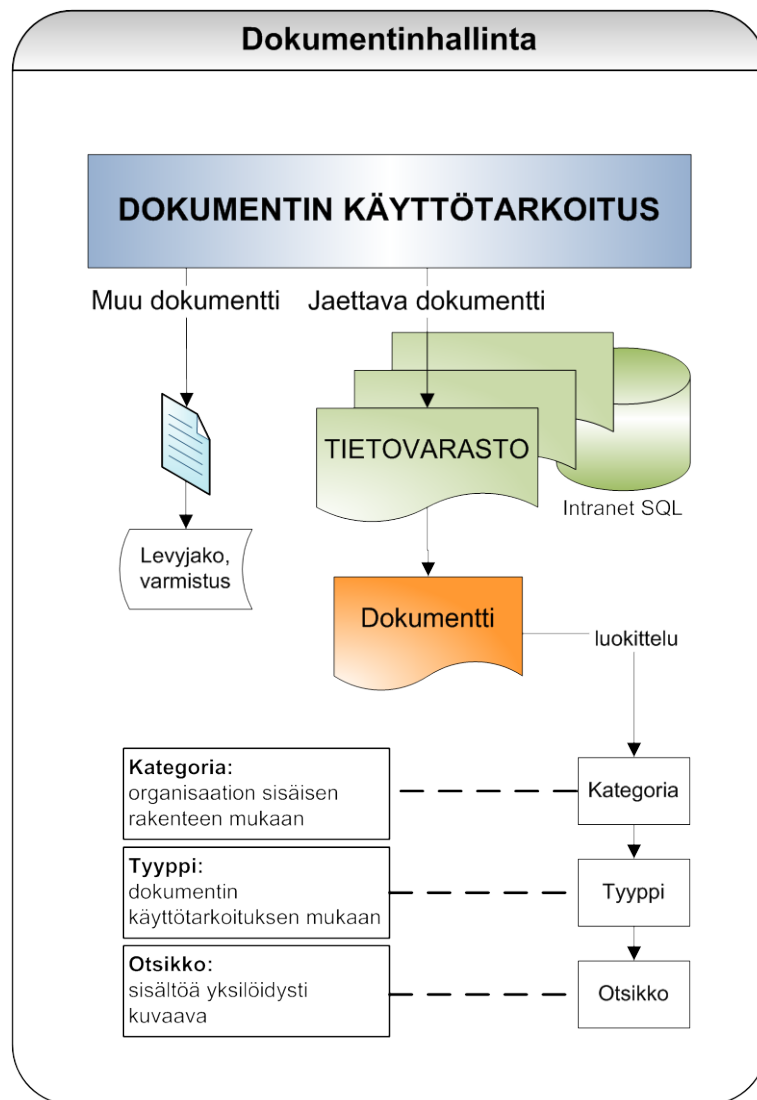
Käyttäjälle ehdotettavina tuloksina hakusanalle ”loma” ovat lomakalenteri, johon merkitään loma, sekä loogisena jatkona lomake poissaoloilmoituksen tekoa varten. Valintojen esittämisessä on käytetty dokumenteista riippumatonta erillistä metatietoa, jolla on annettu nimi ja kuvaus hakutulokselle.

5.5 Ehdotus metatietomalliksi

Tutkittaessa intranetin tiedonhakua ja sisäisen tiedon hallintaa nousi tämän insinööri-työn rajauksessa keskeiseksi kehityskohteeksi metatiedon käyttö yrityksessä. Sitä hyö-

dynnetään paitsi hakutulosten esittämisessä ja hakujen kohdistamisessa, myös erilaisten listausten, näkymien ja hierarkioiden luomisessa dokumentinhallintaan. Tämän vuoksi työn osana syntyi ehdotus yrityksen metatietomalliksi. Intranetin kohdalla metatiedon tuottamiseen osallistuu periaatteessa koko henkilöstö. Siitä johtuen ei mallia voitu määrittellä kovin tiukasti ja laajasti, koska yrityksen sisällä käytetään hyvin erilaisia dokumentteja. Lisäksi käyttäjät voivat helposti kokea dokumentinhallintaan liittyvät rutiinit rasittaviksi, jos he joutuvat liittämään mielestään liikaa pakollisia tietoja dokumenttiin, ennen kuin saavat tallennettua sen järjestelmään.

Metatietomallin kehittämisessä otettiin lähtökohdaksi organisaation rakenne, jonka pohjalta luotiin perusjako dokumenttien luokille. Lisäksi valikoidulle loppukäyttäjryhmälle lähetettiin ehdotus kommentoitavaksi, ja kommenttien perusteella suunnitelmaa tarkennettiin. Näkökulmaan tarvittiin myös painotetusti yrityksen johdon mielipide, sillä metatiedon käytöllä voidaan määrittää myös sellaisia dokumenttien ominaisuuksia, jotka liittyvät käyttöoikeuksiin ja luottamuksellisuuteen. Tämän perusteella pyrittiin hahmottamaan minimivaatimukset metatietomallille. Organisaation rakenteeseen perustuvan luokituksen lisäksi tarvittiin myös muita luokkia, joilla oli looginen yhteys yrityksen toimintaan. Näin muodostettiin *kategoriat dokumenttityypeille*. Kuvassa 20 on esitetty kaaviona dokumentinhallinnan perusjako ja dokumentteihin liitettävät pakolliset ominaisuudet.



Kuva 20. Dokumentinhallinnan luokittelukaavio

Kaavion mukaan jokainen dokumentti on jotain *tyyppiä* ja kuuluu johonkin *kategoriaan*. Lisäksi dokumentilla on *otsikko*, joka kuvaa selväkielisesti dokumentin sisältöä. Ratkaisun lähtökohtana on varmistaa, että järjestelmään ei tallennettaisi irrallisia dokumentteja, vaan ne olisivat aina liitettävissä johonkin asiayhteyteen. Otsikon merkitys korostuu hakutulosten esittämisessä, kuten aiemmissa luvuissa on kuvattu. Nämä kolme metatietokenttää määritettiin pakollisiksi. Niitä ei voitu ottaa käyttöön heti koko intranetissä, koska palvelun käyttöönottoaiheessa järjestelmään tallennettavien vanhojen dokumenttien luokittelu olisi ollut liian haastava tehtävä tässä projektiaikataulussa. Sen vuoksi metatietomalli tuotiin projektityössä hyödynnettävien työtilojen ominaisuuksiksi.

Työtiloissa on vakioidut dokumenttipohjat, jotka saa käyttöön dokumenttikirjaston avulla. Dokumenttia luotaessa ja ensimmäistä kertaa tallennettaessa liitetään metatieto siihen mukaan. Dataa pääsee muokkaamaan suoraan myös dokumenttikirjaston näkymästä. Kuvassa 21 on täyttönäkymä työtilassa käyttöön otetusta dokumenttipohjasta, joka aukeaa Microsoft Wordiin suoraan metatiedon lisäysmahdollisuuden kanssa.

Kuva 21. Metatiedon liittäminen dokumenttiin

Koska dokumentinhallintakäytännöt ovat uusi asia yrityksessä, ei kategorioita ja tyyppiä voitu vielä vahvistaa lopullisesti. Tämän vuoksi mallia kehitettiin ehdotuksena ja ensi vaiheessa päädyttiin taulukossa 7 esitettyihin arvoihin.

Taulukko 7. Metatiedon kategoriat ja tyypit

Kategoria	Tyyppi	
Henkilöstö	Asialista	Raportti
Laatujärjestelmä	Ilmestymisilmoitus	Selvitys
LIKE	Käsikirjoitus	Sopimus
Myynti ja markkinointi	Laatuohje	Suunnitelma
Sisäiset palvelut	Laskelma	Tarjous
T & K	Lomake	Tarjouspyyntö
Talous	Muistio	Tilaus
Tuotanto	Määrittely	Työsääntö
	Ohje	Muu asiakirja
	Pöytäkirja	

Käyttäjäkokemusten ja -palautteen mukaan arvoja voidaan tarvittaessa muuttaa myöhemmin. Poistettavien luokkien kohdalla täytyy silloin tarvittaessa huomioida niihin kuuluvien dokumenttien siirto uuteen luokkaan. Kategorioihin otettiin organisaatiorakenteen ulkopuolelta luokat Laatujärjestelmä, Talous ja Henkilöstö. Ne kuuluvat loogi-

sesti muiden organisaation osien alle, mutta nostettiin omiksi luokikseen ilmeisen dokumenttien tuotantotarpeen vuoksi.

5.6 Tehtyjen muutosten arviointi

Joulukuussa 2009 hakuindeksissä oli yli 5 000 dokumenttia, jotka ovat pääosin Microsoft Office- tai pdf-tiedostoja ja loput ovat intranetin sivustoja ja kuvia. Käyttäjätilastojen perusteella käyttönoton ensimmäisten kuukausien aikana haun käyttö on kasvanut tasaisesti ja marraskuun lopussa keskimääräinen käyttö oli lähes 1 haku/työntekijä päivässä. Tämä korreloi hyvin sen kanssa, että intranetin merkitys yrityksessä on samaan aikaan kasvanut koko ajan ja sen käyttö on tullut henkilöstölle tärkeämmäksi.

Käyttäjätestin perusteella saadut tulokset toimivat hakupalvelun kehittämistä ohjaavina tekijöinä yhdessä muun palautteen ohella. Olennaisena osana testin suunnittelussa oli testitapausten uusimisen mahdollistaminen. Se pystyttiin tekemään tietyissä rajoissa, mutta koska intranetin sisältö kasvoi koko ajan, muuttui myös hakuindeksin sisältö. Siten uusintatellit eivät ole täysin vertailukelpoisia alkuperäisten kanssa, koska hakutulosten saanti on kasvanut merkittävästi tarkasteluajavälillä.

Testitehtäviä kuitenkin käytiin läpi uudestaan kohderyhmän kanssa valikoidusti ja kriittisesti arvioiden. Tulokseksi saatiin merkittävä parannus hakutulosten relevanssissa, kun dokumenttien otsikot nimettiin järkevästi ja kuvaavasti. Sen lopputuloksena dokumentit paitsi sijoittuivat paremmin hakutuloksissa, olivat myös helpommin valittavissa listauksesta. Indeksien laajennus laatu järjestelmän sisältöön toi haun kautta käytettäväksi keskeistä sisältöä. Mahdollisuus rajata hakutuloksia koettiin tärkeäksi parannukseksi. Ongelma suomen kielen yhdyssanojen käytössä voitiin kiertää ohjeistamalla, mutta palautteessa toivottiin järjestelmään korjausta tältä osin.

Kun käytettiin samoja hakusanoja kuin aiemmin, löytyi relevantti tulos uusintakokeissa aina ensimmäiseltä hakutulossivulta. Sijoituksen keskiarvoa ei ollut tässä tilanteessa mielekäästä laskea, koska tuloksia palautui paljon enemmän kuin testivaiheessa, eikä sopivaa painotuskerrointa voitu määrittää. Tiedon löytyminen sen sijaan nopeutui merkittävästi ja mediaani nousi kaksi luokkaa ylöspäin. Tätä tulosta täytyy tosin tarkastella

kriittisesti siitä näkökulmasta, että käyttäjät ovat oppineet lukemaan hakutuloksia ja tiesivät paremmin, mitä hakea ja odottaa. Joka tapauksessa vapaassa käyttäjäpalautteessa pidettiin hyvänä haun teknisen toimintanopeuden lisäksi parantunutta tiedon löydettävyyttä myös ajallisesti. Tämä on selvässä korrelaatioissa hakutulosten kasvaneeseen relevanssiin ja parantuneeseen esitystapaan.

Metatiedon hyödyntämisen tarkka arviointi jää pääosin tämän työn aikataulun ja resursien ulkopuolelle, sillä kokonaisuuden hahmottaminen vaatii pidemmän aikavälin seuranta ja kehitystyötä. Metatiedon käyttö dokumenttien otsikoinnissa ja sen suhde hakutuloksien esittämisessä oli kuitenkin opettava tekijä ja sai loppukäyttäjät ymmärtämään konkreettisemmin intranetin ja yrityksen sisäisen tiedon luokittelun tarvetta.

6 Yhteenveto

Insinööriyössä tutkittiin Rakennustieto Oy:n intranetin hakutoimintoa yrityksen tietämyksenhallinnan kehittämisen näkökulmasta. Kehitystyö pohjautui käyttäjätutkimukseen, jonka tuloksia tuettiin hakujärjestelmän ominaisuuksien heuristisella arvioinnilla. Testit teetettiin todellisilla loppukäyttäjillä, ja sillä pyrittiin löytämään järjestelmän käytettävyydessä ja toimivuudessa esiintyvät puutteet. Tärkeimpänä mittarina toimi käyttäjän tarvitseman relevantin tiedon nopea ja helppo löydettävyys.

Työn teoreettinen tausta perustui yleiseen tiedonhaun teoriaan (information retrieval) ja informaatiotutkimuksessa käytettäviin tiedon luokitusmenetelmiin. Näitä teorioita tuettiin semanttisen webin ja tietämyksenhallinnan käsitteistöllä. Keskeisenä lähtökohtana oli tiedon merkityksen määrittely relevanssin käsitteen avulla. Työn kannalta teoriasta parhaiten hyödynnettävissä oli metatiedon käyttö yrityksen sisäisen tiedon luokittelussa ja edelleen hakutulosten relevanssin kehittämisessä. Käytetyn hakukoneen avainsanalyyttien muokkaaminen antoi mahdollisuuden esittää hakutuloksia, jotka pohjautuvat periaatteeltaan ontologiseen määrittelyyn. Yrityksen dokumentinhallinnan tueksi kehitetty ehdotus metatietomallista perustuu lähtökohdaltaan ontologiseen mallintamiseen. Tämän työn ja käytetyn järjestelmän puitteissa sitä ei kuitenkaan voitu viedä tässä vaiheessa pidemmälle.

Hakutoiminto otettiin käyttöön intranetin toimituksen mukana tullessa perusominaisuutena. Jo sellaisenaan se koettiin selvänä parannuksena aikaisempaan verrattuna. Tutkimuksen perusteella tehdyt kehitystoimenpiteet osoittivat, että intranetin haku toimii ja sitä voidaan kehittää jokapäiväiseen työhön kiinteästi kuuluvana työkaluna. Tiedonhaakuun käytetyn ajan väheneminen vapauttaa työaikaa muihin tehtäviin, ja siten se mahdollistaa innovatiivisemmän ja tehokkaamman työskentelyn. Tässä työssä ajankäytön paranemista arvioitiin ainoastaan hakukoneen kehittämisestä saaduilla hyödyillä. Työnantajan kannalta vielä olennaisempaa on se, miten paljon tehokkaampaa tiedonhaku on hyvin toimivan haun kanssa kuin kokonaan ilman sitä.

Työn tarkastelunäkökulmana oli tietämyksenhallinnan kehittäminen yrityksessä. Hakutoiminnon kehitys auttoi tässä tavoitteessa tarjoamalla tehokkaamman ja kontrol-

loidumman tiedon tavoitettavuuden. Lisäksi haku tarjoaa tuloslistauksen kautta tietoa muista dokumenteista, jotka ovat sisällöltään vastaavia tai lähellä haettua tietoa. Tämä auttoi loppukäyttäjää muodostamaan laajemman käsityksen asiaan liittyvästä kokonaisuudesta. Tämän työn kannalta oli olennaista kehittää tulosten relevanssia siten, että tuloslistauksen relevanssi mahdollisti oheistiedon löytymisen. Lisäksi käyttäjälle tarjoutui mahdollisuus löytää tuloksista samaan asiaan liittyen täysin toisia aihealueita, kuin mihin haettu asia liittyy. Sitä voidaan jatkossa hyödyntää tiedon integraatiossa ja sitä kautta tietämyksen kehittämisessä.

Insinööriyön tekeminen sujui pääosin ongelmattomasti. Hakutoiminnon testaamiseksi oli olennaista, että sisältömassaa on riittävästi. Sen kerryttämiseksi oli käytettävissä oma aliprojekti. Järjestelmätoimitukseen kuulunut takuu ja ylläpitosopimus rajoittivat kehitystyössä tehtyjä kokeiluja tietyissä tapauksissa. Hakutulosten sijoitusten parantamiseksi ei voitu käyttää kaikkia keinoja, koska ne olisivat vaatineet järjestelmän ohjelmointirajapinnan käyttöä, mikä oli ristiriidassa toimitussopimuksen kanssa. Testitapausten uusimisen validiteetin kannalta olisi ollut hyvä, jos intranetin sisältö olisi ollut jo suhteellisen vakiintuneella tasolla ja jos muutostyöt olisi pystytty ennakoiden tekemään mahdollisimman pian testin jälkeen. Uusintatestaus olisi silloin voitu tehdä mahdollisimman samanlaisessa ympäristössä.

Tiedonhaun kehittäminen koettiin yrityksessä tärkeänä asiana. Kokonaisuuden kehittämiseksi jatkossa on tärkeää saada yrityksen sisäinen dokumentinhallinta mahdollisimman kattavasti intranetin kautta toimivaksi. Samalla se voidaan liittää metatietomalliin, joka tarjoaa mahdollisuuden luokitella ja tehostaa tiedonkäyttöä yrityksessä. Ongelmana kehitysprosessissa on käyttäjien sitouttaminen metatiedon käyttöön ja tuottamiseen. Se pitäisi tehdä riittävän automaattiseksi, ja toisaalta käyttäjiä on pyrittävä motivoimaan tiedontuotantoon sen hyötyjä korostamalla. Parhaiten se onnistuu käyttötilanteissa tapahtuvana syy-yhteyksien ymmärtämisenä, kuten dokumentin otsikon vaikutus hakutulokseen tässä työssä osoitti. Lisäksi valmiudessa oli kehitystyön yhteydessä havaittavissa eroja eri ikäryhmien välillä. Perusajatuksena on saada käyttäjät kuvailemaan aineistojaan itse mahdollisimman varhaisessa vaiheessa tiedon tuotantoketjua.

Haun kehittäminen jatkossa perustuu käyttäjäpalautteen ja käyttötilastojen seurantaan sekä uusien tekniikoiden hyödyntämiseen. Yrityksen sisäisesti on mahdollista analysoida käyttötapauksia ja -tilanteita, joiden mukaan voidaan ryhmitellä tiedon käyttötarpeita. Tämä voi toimia lähtökohtana yrityksen sisäisille ontologioille, jossa tiedonhaku voidaan ohjata työntekijän profiilin mukaan mahdollisimman tehokkaasti. Kokonaisuutena tiedonhaku osana yrityksen tietämyksenhallintaa on hyvä nähdä menestystekijänä yrityksen toiminnassa. Mikäli koko prosessia voidaan mitata ja kohdistaa tulokset yrityksen toiminnan kehittämiseen, se voi toimia myös kilpailuetuna.

Lähteet

1. Wong, Kuan Yew. Critical success factors for implementing knowledge management in small and medium enterprises. Teoksessa *An Introduction to Knowledge Management ...an Emerald Guide*. Bradford, UK: Emerald Group Publishing, 2005.
2. Ståhle, Pirjo & Grönroos, Mauri. Knowledge management – tietopääoma yrityksen kilpailutekijänä. Porvoo: WSOY, 1999.
3. Kuivalahti, Tatu & Luukkonen, Jussi. Intra. Helsinki: Inforviestintä, 2003.
4. Grönroos, Mauri. Mahdollisuuden aika – kohti virtuaalista organisaatiota. Tampere: Transatlanta, 2003.
5. Feldman, Susan. The high cost of not finding information. *KM World magazine: KM in Practice*. March 2004 [Volume 13, Issue 3]. (WWW-dokumentti.) <<http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=9534>>. 1.3.2004. Luettu 27.4.2009.
6. Niiniluoto, Ilkka. Informaatio, tieto ja yhteiskunta. Helsinki: Valtion painatuskeskus, 1988.
7. Brelade, Sue. *Practical Guide to Knowledge Management*. London, UK: Thorogood, 2003.
8. Levene, Mark. *An Introduction to Search Engines and Web Navigation*. Essex, UK: Pearson Education, 2006.
9. Losee, R. M. Term dependence: A basis for Luhn and Zipf models. *Journal of the American Society for information Science and Technology (JASIST)*, 2001. Vol. 52, s. 1019–1025.
10. Witten, Ian H., Gori, Marco & Numerico, Teresa. *Web Dragons: Inside the Myths of Search Engine Technology*. San Francisco, CA: Morgan Kauffman, 2006.
11. Guha, R., McCool R., and Miller E. Semantic search. *Proceedings of the 12th International Conference on World Wide Web*, May 20–24, 2003. Budapest, Hungary. S. 700–709.
12. Intranet. (WWW-dokumentti.) Tekniikan sanastokeskus. <<http://www.tsk.fi/tepa/>>. 25.10.1999. Luettu 23.11.2010.
13. Collins, Heidi. *Enterprise Knowledge Portals: Next Generation Portal Solutions for Dynamic Information Access, Better Decision Making and Maximum Results*. Sa-

- ranac Lake, NY: Amacom, 2002.
14. Davies, John (editor). Semantic Knowledge Management. Bradford, UK: Emerald Group Publishing, 2005.
 15. Ojala, Leenamajja & Pöysti, Kaija. Wikimaniaa yrityksiin: Yritys 2.0 tuottamaan. Helsinki: WSOYpro, 2008.
 16. Yu, Liyang. Introduction to the Semantic Web and Semantic Web Services. Boca Raton, FL: Taylor & Francis Group, 2007.
 17. Valtonen, Marja. Huono hakutoiminto pahin virhe. IT-Viikko 19.5.2005, s. 6.
 18. Haasio, Ari, Piukkula, Juha & Ekholm, Kai. Tietoverkot ja kirjastot. Helsinki: BTJ Kirjastopalvelut, 2002.
 19. Brin, Sergey & Page, Lawrence. The Anatomy of a Large-Scale Hypertextual Web Search Engine. (WWW-dokumentti.)
<<http://infolab.stanford.edu/~backrub/google.html>>. Luettu 14.1.2010.
 20. Eronen, Riitta. Ennätysten kieli. (WWW-dokumentti.)
<<http://www.kotus.fi/index.phtml?s=2747>>. 25.4.1997. Luettu 23.12.2009.
 21. Karlsson, Fred. Johdatusta yleiseen kielitieteeseen. Vaasa: Oy Gaudeamus Ab, 1982.
 22. Tisseghem, Patrick & Fastrup, Lars. Inside the index and search engines. Redmond, Washington: Microsoft Press, 2008.
 23. Antoniou, Grigoris & van Harmelen, Frank. A Semantic Web Primer. MIT, 2004.
 24. Berners-Lee, Tim, Hendler, James & Lassila, Ora. The Semantic Web. Scientific American. May 2001. (WWW-dokumentti.)
<<http://www.scientificamerican.com/article.cfm?id=the-semantic-web>>. 17.5.2001. Luettu 20.3.2009.
 25. Bhavani, Thuraisingham. Building Trustworthy Semantic Webs. Boca Raton, FL: Auerbach Publications, 2008.
 26. Hyvönen, Eero, Seppälä, Katri, Viljanen, Kim & Frosterus, Matias. Yleinen suomalainen ontologia YSO – kohti suomalaista semanttista webiä (General Finnish Ontology YSO – Towards the Finnish Semantic Web). Tietolinja, 2007.

27. Understanding Metadata. Bethesda, MD: NISO Press National Information Standards Organization, 2004.
28. Metadata Basics. Dublin Core Metadata Initiative. (WWW-dokumentti.) <<http://dublincore.org/metadata-basics/>>. Luettu 14.1.2010.
29. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation. (WWW-dokumentti.) <<http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/#section-data-model>>. 10.2.2004. Luettu 1.1.2010.
30. Champin, Pierre-Antoine. RDF Tutorial. (WWW-dokumentti.) <<http://www710.univ-lyon1.fr/~champin/rdf-tutorial/rdf-tutorial.pdf>>. 5.4.2001. Luettu 1.1.2010.
31. OWL Web Ontology Language Overview. W3C Recommendation. (WWW-dokumentti.) <<http://www.w3.org/TR/owl-features/>>. 10.2.2004. Luettu 30.11.2009.
32. Sinkkonen, Irmeli, Kuoppala, Hannu, Parkkinen, Jarmo & Vastamäki, Raino. Käytettävyyden psykologia. Helsinki: Edita IT Press, 2006.
33. Nielsen, Jacob. Usability Engineering. Boston: Academic Press Professional, 1993.
34. Schamber, L., Eisenberg, M. & Nilan, M. A re-examination of relevance: toward a dynamic, situational definition. Information Processing and Management, 1990. Vol. 26, s. 755–776.
35. Nielsen, J. Usability for the masses. Journal of Usability Studies, 2005. Vol. 1, s. 2–3.
36. Nielsen, Jacob. Ten Usability Heuristics. (WWW-dokumentti.) <http://www.useit.com/papers/heuristic/heuristic_list.html>. Luettu 4.1.2010.

Liite 1: Kyselylomake

Intranetin hakutoiminnon testaus

Tällä kyselyllä tutkitaan Rakennustiedon intranetin (osoitteessa <http://intra/>) vapaaseen sanahakuun perustuvan hakukoneen toimivuutta. Kyselyn lopputuloksia käytetään yrityksen sisäisen tiedonhaun kehittämiseen ja se on osa Mikko Sjöbergin insinööriyötä.

Kysely perustuu verkkopalvelun käytettävyytestauksen menetelmiin ja se käsittää 5 kpl hakutehtäviä, joiden suorittamisen vaiheet testihenkilö dokumentoi mahdollisimman tarkkaan. Testaajalla on lopuksi mahdollisuus vapaaseen kommentointiin. Lisäksi testitulosten perusteella voidaan järjestää tarvittaessa lisähaastattelu ja testitapausten vertailu järjestelmään tehtyjä muutoksien toimivuuden tarkastamiseksi. Vastattuasi kyselyn kaikkiin kohtiin lähetä lomake painamalla lopussa olevaa "Lähetä"-painiketta. Annan tarvittaessa lisätietoja kyselystä. Testiin kuluu aikaa alle n. ½ h ja tulokset käsitellään luottamuksellisesti.

Testin kulku

Testissä tehdään ensin 5 kpl hakutehtäviä, joista 4 on ennalta osoitettuja ja 1 on vapaa- valintainen, käyttäjän omaan tarpeeseen perustuva tehtävä. Tehtävissä pyritään löytämään Rakennustiedon intranetistä tieto tai dokumentti ja samalla merkitään muistiin, kuinka monta valintaa käyttäjä joutuu tekemään tyydyttävään lopputulokseen päästäkseen.

Hakutehtävien suoritus

1. Avaa intranet Internet Explorer -selaimen osoitteessa <http://intra/>
2. Kirjoita kaikki testissä käyttämäsi hakusanat selainikkunan yläreunassa olevaan hakukenttään ja paina "Hae"-painiketta. Hakutulokset palautuvat listauksena selainikkunaan ja yksittäinen dokumentti aukeaa linkistä kuten yleisissä www-hakukoneissa
3. Käytä tehtävien tulosten kirjaamisessa **tässä lomakkeessa** olevia kenttiä ja niiden yhteydessä **annettuja ohjeita**
4. Muista kirjata ylös kaikki käyttämäsi hakusanat, myös sellaiset, joita täydennät muokkaamalla tai lisäämällä jotain jo käytettyyn hakusanaan. Esim. hakusanat esityslista ja kokouksen esityslista ovat kaksi eri hakusanaa ja ne tulee kirjata erikseen
5. Lähetä lopuksi täytetty lomake sähköpostilla "Lähetä"-painiketta käyttäen

1. Tehtävä

Olet järjestämässä kokousta ja tarvitset tarjoilun kokoukseen. Tiedät, miten tarjoilu tilataan, mutta et muista eri vaihtoehtoja. Etsi intranetistä, millaisia kokoustarjoiluvaihtoehtoja on olemassa.

<p>Käytetyt hakusanat ja niiden yhdistelmät. Kirjoita <u>viimeiseksi</u> paras hakusana (erota hakusanat pilkulla)</p>	
--	--

Kuinka mones hakutulos tuotti tyydyttävän ratkaisun?

Tulos esiintyi sivulla nro ja oli sivulla linkki nro (ylhäältä laskettuna)

Mikä otsikko tai muu tekijä johdatti valitsemaan oikean dokumentin?

<p>Merkitse valintapäätökseen vaikuttanut teksti tähän (voit myös kopioida selaimesta)</p>	
--	--

Montako turhaa valintaa (=avaamaan dokumentin) jouduit tekemään, ennen kuin päädyit oikeaan dokumenttiin?

Ylimääräisiä valintoja oli

Löytyikö haettu tieto tai dokumentti?

Kyllä Ei

Arvio hakutehtävään käytetystä ajasta:

alle 15 s 15 - 30 s 30 - 60 s 1 - 2 min yli 2 min

Lopuksi vielä arvioi hakutehtävän onnistumista kokonaisuutena asteikolla 5 = onnistunut 1= epäonnistunut

5 4 3 2 1

2. Tehtävä

Henkilöstöhallinto haluaa kesälomasi jälkeen tiedon käyttämistäsi lomapäivistä. Etsi intranetistä lomake, jolla voit ilmoittaa tiedon heille.

<p>Käytetyt hakusanat ja niiden yhdistelmät. Kirjoita <u>viimeiseksi</u> paras hakusana (erota hakusanat pilkulla)</p>	
--	--

Kuinka mones hakutulos tuotti tyydyttävän ratkaisun?

Tulos esiintyi sivulla nro ja oli sivulla linkki nro (ylhäältä laskettuna)

Mikä otsikko tai muu tekijä johdatti valitsemaan oikean dokumentin?

<p>Merkitse valintapäätökseen vaikuttanut teksti tähän (voit myös kopioida selaimesta)</p>	
--	--

Montako turhaa valintaa (=avaamaan dokumentin) jouduit tekemään, ennen kuin päädyit oikeaan dokumenttiin?

Ylimääräisiä valintoja oli

Löytyikö haettu tieto tai dokumentti?

Kyllä Ei

Arvio hakutehtävään käytetystä ajasta:

alle 15 s 15 - 30 s 30 - 60 s 1 - 2 min yli 2 min

Lopuksi vielä arvioi hakutehtävän onnistumista kokonaisuutena asteikolla 5 = onnistunut 1= epäonnistunut

5 4 3 2 1

3. Tehtävä

Asiakaspuhelu päättyy sattumalta sinulle. Asiakas tarvitsee tietoa vanhan rakennuksen purkamisesta. Muistat lukeneesi äskettäin asiasta intrasta ja päätät hakea kyseisen tuotteen tiedot hakukoneen avulla ja auttaa asiakasta.

Käytetyt hakusanat ja niiden yhdistelmät. Kirjoita viimeiseksi paras hakusana (erota hakusanat pilkulla)	
---	--

Kuinka mones hakutulos tuotti tyydyttävän ratkaisun?

Tulos esiintyi sivulla nro ja oli sivulla linkki nro (ylhäältä laskettuna)

Mikä otsikko tai muu tekijä johdatti valitsemaan oikean dokumentin?

Merkitse valintapäätökseen vaikuttanut teksti tähän (voit myös kopioida selaimesta)	
--	--

Montako turhaa valintaa (=avaamaan dokumentin) jouduit tekemään, ennen kuin päädyit oikeaan dokumenttiin?

Ylimääräisiä valintoja oli

Löytyikö haettu tieto tai dokumentti?

Kyllä Ei

Arvio hakutehtävään käytetystä ajasta:

alle 15 s 15 - 30 s 30 - 60 s 1 - 2 min yli 2 min

Lopuksi vielä arvioi hakutehtävän onnistumista kokonaisuutena asteikolla 5 = onnistunut 1= epäonnistunut

5 4 3 2 1

4. Tehtävä

Olet mukana messutapahtumaa valmistelevassa työryhmässä ja tehtävänäsi on selvittää messuosaston suunnitteluun ja rakentamiseen sopivia yrityksiä. Et tiedä asiasta entuudestaan mitään ja päätät hakea tietoa intrasta.

<p>Käytetyt hakusanat ja niiden yhdistelmät. Kirjoita <u>viimeiseksi</u> paras hakusana (erota hakusanat pilkulla)</p>	
--	--

Kuinka mones hakutulos tuotti tyydyttävän ratkaisun?

Tulos esiintyi sivulla nro ja oli sivulla linkki nro (ylhäältä laskettuna)

Mikä otsikko tai muu tekijä johdatti valitsemaan oikean dokumentin?

<p>Merkitse valintapäätökseen vaikuttanut teksti tähän (voit myös kopioida selaimesta)</p>	
--	--

Montako turhaa valintaa (=avaamaan dokumentin) jouduit tekemään, ennen kuin päädyit oikeaan dokumenttiin?

Ylimääräisiä valintoja oli

Löytyikö haettu tieto tai dokumentti?

Kyllä Ei

Arvio hakutehtävään käytetystä ajasta:

alle 15 s 15 - 30 s 30 - 60 s 1 - 2 min yli 2 min

Lopuksi vielä arvioi hakutehtävän onnistumista kokonaisuutena asteikolla 5 = onnistunut 1= epäonnistunut

5 4 3 2 1

5. Tehtävä: hae vapaavalintainen dokumentti intrasta

Kirjota tai kuvaa tähän tieto tai dokumentti, jota tarvitset usein.	
---	--

Käytetyt hakusanat ja niiden yhdistelmät. Kirjoita <u>viimeiseksi</u> paras hakusana (erota hakusanat pilkulla)	
---	--

Kuinka mones hakutulos tuotti tyydyttävän ratkaisun?

Tulos esiintyi sivulla nro ja oli sivulla linkki nro (ylhäältä laskettuna)

Mikä otsikko tai muu tekijä johdatti valitsemaan oikean dokumentin?

Merkitse valintapäätökseen vaikuttanut teksti tähän (voit myös kopioida selaimesta)	
---	--

Montako turhaa valintaa (=avaamaan dokumentin) jouduit tekemään, ennen kuin päädyit oikeaan dokumenttiin?

Ylimääräisiä valintoja oli

Löytyikö haettu tieto tai dokumentti?

Kyllä Ei

Arvio hakutehtävään käytetystä ajasta:

alle 15 s 15 - 30 s 30 - 60 s 1 - 2 min yli 2 min

Lopuksi vielä arvioi hakutehtävän onnistumista kokonaisuutena asteikolla 5 = onnistunut 1= epäonnistunut

5 4 3 2 1

Vapaa kommentointi- ja arviointimahdollisuus: mikä toimi ja mikä ei, toimiko haku nopeasti, mitä lisätoiveita ja parannusideoita sait testauksen aikana?

--

Lähetä lopuksi täytetty lomake ensimmäisellä tai viimeisellä sivulla olevalla "Lähetä" -painikkeella. Voit halutessasi toimittaa täytetyn lomakkeen myös paperitulosteena.