Vaatimusdokumentti

Metaxa

Helsinki 11.12.2005
Ohjelmistotuotantoprojekti
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Kurssi

581260 Ohjelmistotuotantoprojekti (6 ov)

Projektiryhmä

Väinö Ala-Härkönen
Reima Halmetoja
Antti Laitinen
Kalle Pyykkönen
Oskari Saarekas
Tuomas Tanner
Juuso Vanonen

Asiakas

Olli Niinivaara

Johtoryhmä

Juha Taina
Joni Salmi

Kotisivu

`http://www.cs.helsinki.fi/group/metadata/`

Versiohistoria

13.9.2005 Versio 1
14.9.2005 Versio 2
20.9.2005 Versio 3
29.9.2005 Versio 4
10.10.2005 Versio 5
11.10.2005 Versio 6
11.10.2005 Versio 6.1 (Väinö oikolukenut ja kommentoinut)
13.10.2005 Versio 6.2 (TR-korjaukset)
14.10.2005 Versio 6.3 (TR-korjauksia, ym.)
17.10.2005 Versio 7
18.10.2005 Versio 7.1 (TR-korjauksia)
20.10.2005 Versio 8 (Transformaatiosäännöt & heuristiikat uutta TR varten)
21.10.2005 Versio 9 Lopullinen hyväksytettävä versio
21.10.2005 Versio 10 Jäädytetty (asiakkaan hyväksymä)
18.11.2005 Versio 11 Sulatettu (korjattu nimiformaattin transformaatio)
22.11.2005 Versio 12 Karsittu ja muutettu selausvaiheen vaatimuksia
28.11.2005 Versio 13 Muokattu transformoinnissa tapahtuvaa kanonisointia
08.12.2005 Versio 14 Karsittu vaatimuksia, tarkistettu
11.12.2005 Versio 15 Lopulliset korjaukset, puhtaaksikirjoitus

Sisältö

1. Johdanto

Metadata on Helsingin yliopiston tietojenkäsittelytieteen laitoksella toteutettava ohjelmistotuotantoprojekti. Projektissa toteutetaan metadatan hallintatyökalu laitoksen tutkijan Olli Niinivaaran käyttöön. Asiakas haluaa tutkia tietojenkäsittelytieteen artikkeleista hajautetuissa paikoissa sijaitsevaa bibliografista metadataa.

Asiakas haluaa tarkastella tätä hajanaisessa muodossa olevaa metadataa yhdistetyssä muodossa siten, ettei metadatan alkuperäinen muoto vaikuta tarkasteluun. Tarkastelu mahdollistetaan muokkaamalla eri formaateissa olevaa metadataa ohjelmiston käyttämään yhteiseen formaattiin. Yhteisessä formaatissa olevasta metadatasta koostetaan resursseja, jotka kuvaavat kirjoittajia, dokumentteja, julkaisijoita ja muita tietojenkäsittelytieteen artikkeleihin liittyviä tosielämän resursseja ja niiden välisiä suhteita. Resurssiverkon avulla voidaan käsitellystä tekstimuotoisesta metadatasta hahmottaa visuaalinen kokonaisuus.

Resurssiverkkosta tehdään tarkastelussa poimintoja, joissa tarkasteltavia resursseja rajataan tietyillä kriteereillä. Resurssien välisiä yhteyksiä voidaan hallinnoida ja näihin yhteyksiin voidaan kiinnittää painotuksia. Tarkastelun tulokset voidaan tallentaa myöhemmin muilla työkaluilla tapahtuvaa tiedon louhintaa varten. Resurssiverkossa tapahtuva poiminta mahdollistaa tekstimuotoisen metadatan muokkaamista intuitiivisemman ja tuottavamman menetelmän jatkokäsittelyssä tarvittavan datan tuottamiseen.

Luvussa 2 esitellään järjestelmään liittyvät määritelmät, järjestelmän yleiskuva ja liittymät muihin järjestelmiin. Asiakkaan vaatimukset järjestelmälle on esitetty luvussa 3. Vaatimusten perusteella määritellyt järjestelmän toiminnot ja niiden yhteiset määritykset on esitelty luvussa 4.

2. Järjestelmän yleiskuvaus

2.1. Määritelmät

Metadata

Tieto, joka kuvaa muuta tietoa.

Attribuutti (attribute)

Metadatatietueeseen liittyvä ominaisuus & arvo -pari, joka määrittelee yhden tietueen ominaisuuden.

Metadatatietue (metadata record)

Attribuuteista koostuva dokumentin kuvaus joka määrittää yhden tai useamman resurssin.

Raakadata (raw data)

Palvelimella tai paikallisessa tietovarastossa sijaitseva metadata jota ei ole jäsennetty eikä muokattu tässä ohjelmassa.

Lähde (source)

Raakadataa sisältävä tietovarasto kuten palvelin tai tiedostopolku.

Atomilause (atomic clause)

Transformoinnin tuloksena saatu tietueen osa, jota ei voi enää jakaa pienempiin osiin. Toteutettavassa ohjelmistossa atomilause koostuu viittauksesta tietueeseen, tietueen ominaisuudesta ja tämän ominaisuuden arvosta.

Dublin Core

Informaatioresurssien kuvaamiseen tarkoitettu standardoitu, yksinkertainen ja laajennettavissa oleva metadata-elementtijoukko, http://dublincore.org/index.shtml

OAI-harvesteri (OAI-harvester)

Ohjelmakomponentti jonka avulla voidaan noutaa metadataa OAI-PMH-protokollan yli.

Pikaformaatti (quick format)

Asiakkaan ohjelmistoa varten määrittelemä yksinkertainen metadatan esitysmuoto. Katso tarkempi määrittely kohdassa "Pikaformaatin määrittely".

DBLP

"Digital Bibliography & Library Project", Internetissä sijaitseva kokoelma joka tarjoaa bibliografista informaatiota tietojenkäsittelytieteen artikkeleista. Lyhennettä käytetään myös kokoelman tiedostoformaatista. Katso http://dblp.uni-trier.de/.

BibTeX

LaTeX-ohjelmistossa käytetty bibliografia-metadataformaatti. Katso http://www.ecst.csuchico.edu/~jacobsd/bib/formats/bibtex.html.

CS-BibTex tietokanta (CS-BibTex database)

Internetissä sijaitseva kokoelma BibTeX-formaatissa olevia bibliografioita tietojenkäsittelytieteen kirjallisuudesta lukuisista lähteistä. Katso http://liinwww.ira.uka.de/bibliography/index.html.

Transformointi (transformation)

Prosessi jossa eri tyyppisiä metadataformaatteja muunnetaan yhteiseen atomilauseformaattiin.

Integrointi (integration)

Prosessi jossa luodaan resurssiverkko atomilauseformaatissa olevasta datasta.

Harvesteri (harvester)

Ohjelmiston osa, joka lataa palvelimilta metadataa soveltuvan tiedonsiirtoprotokollan yli

Harvestointi (harvesting)

Metadatan kerääminen Internetissä sijaitsevalta palvelimelta tai palvelimilta jonkin tiedonsiirtoprotokollan yli.

OAI-PMH

OAI:n määrittelemä protokolla metadatan harvestointiin. Protokollan spesifikaatio sijaitsee www-osoitteessa: http://www.openarchives.org/OAI/openarchivesprotocol.html

Resurssi (resource)

Todellisen maailman kohdetta kuvaava olio joilla kuvataan integroidun metadatan informaatiosisältö. Resursseihin liittyy yksikäsitteinen tunniste, joka muodostetaan resurssin nimestä ja juoksevasta numerosta, sekä hakusanoja. Kuhunkin resurssiin voi tämän lisäksi liittyä merkityksiä ja kuhunkin toimijan ja dokumentin väliseen yhteyteen liittyy jokin rooli.

Toimija (actor)

Resurssityyppi joka kuvaa yksittäistä henkilöä tai tahoa. Kuhunkin toimijaan voi liittyä useita kanavia ja dokumentteja.

Rooli (role)

Resurssityyppi joka kuvaa toimijan ja muuntyyppisen resurssin välistä yhteyttä.

Dokumentti (document)

Resurssityyppi johon liittyy tieto jonkin reaalimaailman dokumentin esitystyypistä, kielestä sekä formaatista ja myös kyseessä olevan dokumentin kuvaus.

Merkitys (meaning)

Reaalilukuarvoinen paino joka liitetään johonkin merkitysyhteyteen tai resurssiin.

Kanava (channel)

Resurssityyppi joka kuva kanavaa jonka kautta jokin dokumentti on julkaistu. Esimerkiksi jonkin artikkelin julkaisukanava voi olla lehti, tms. julkaisu.

Resurssiverkko (resource network)

Resurssien ja niiden välisten yhteyksien muodostama suunnattu, ei-painotettu verkko.

Yhteysverkko (meaning network)

Resurssiverkon aliverkko

Merkitysyhteys (meaning relation)

Kaksi erillistä resurssia yhdistävä yhteys joka painotetaan reaaliluvulla.

Merkitysverkko (meaning network)

Resurssien joihin liittyy merkitys ja niitä yhdistävien merkitysyhteyksien muodostama suunnattu, painotettu verkko

Pajek

Ohjelmisto suurten verkostojen analysoimiseen, katso http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Pajek -listaformaatti

Pajek -ohjelmiston käyttämä verkkojenkuvausformaatti. Tiedostomuoto määritelty käyttöoppaan kohdassa 5.3. http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.pdf

3. Vaatimukset järjestelmälle

3.1. Toiminnalliset vaatimukset

K1.1 Valmiin raakadatan tuonti

Käyttäjä voi tuoda ohjelmaan raakadataa, joka transformoidaan yhteiseen atomilausemuotoon (transformoinnin määrittely luku 5)

Prioriteetti: 1

K1.1.1 XML Dublin Core harvestointi

Käyttäjä voi tuoda lähteistä XML Dublin Core 1.1 Simple -muotoista raakadataa OAI-PMH 2.0:n yli.

Prioriteetti: 1

K1.1.2 Pajek-tiedostojen tuonti

Käyttäjä voi tuoda Pajek-listaformaatti -muotoista raakadataa tiedostoista.

Prioriteetti: ei toteuteta

K1.1.3 DBLP-tiedostojen tuonti

Käyttäjä voi tuoda raakadataa suoraan DBLP-muotoisista tiedostoista.

Prioriteetti: ei toteuteta

K1.1.4 BibTeX metadatan tuonti

Käyttäjä voi tuoda BibTeX -muotoista (versio 0.99b, katso http://bibliographic.openoffice.org/btxdoc.html) raakadataa.

Prioriteetti: ei toteuteta

K1.1.4.1 BibTeX tiedostojen tuonti

Käyttäjä voi tuoda raakadataa suoraan BibTeX-muotoisista tiedostoista.

Prioriteetti: ei toteuteta

K1.1.4.2 BibTeX tietokantaluku

Käyttäjä voi ladata BibTeX-muotoista raakadataa CS-BibTex-tietokannasta.
Prioriteetti: ei toteuteta

K1.1.5 oai_citeseer-muotoisten tiedostojen tuonti

Käyttäjä voi tuoda oai_citeseer-muotoista (http://copper.ist.psu.edu/oai/oai_citeseer/) raakadataa tiedostosta.
Prioriteetti: 1

K1.1.6 XML Dublin Core-muotoisten tiedostojen tuonti

Käyttäjä voi tuoda XML Dublin Core 1.1 Simple -muotoista raakadataa tiedostosta.
Prioriteetti: 1

K1.2 Itse tehdyn metadatan tuonti

Käyttäjä voi tuoda ohjelmaan itse tekemäänsä metadataa, joka transformoidaan yhteiseen atomilausemuotoon (transformoinnin määrittely: kts. Luku 5)
Prioriteetti: 1

K1.2.1 Pikaformaatti-tiedostojen tuonti

Käyttäjä voi tuoda raakadataa yksinkertaisessa pikaformaatissa olevista tiedostoista (pikaformaatin määrittely liitteessä 3).
Prioriteetti: 1

K1.3 Metadatan säilytys

Käyttäjän tuoma metadata säilyy järjestelmässä.
Prioriteetti: 2

K1.3.1 Raakadatan säilytys

Raakadata säilyy järjestelmässä samassa formaatissa ja semanttisesti samansisältöisenä.
Prioriteetti: 2

K1.3.2 Atomilauseiden säilytys

Raakadata säilyy järjestelmässä sen atomilausemuodossa.

Prioriteetti: 3

K1.4 Lähteiden päivitys

Käyttäjä voi päivittää lähteitä vastaavan raakadatan riippumatta datalähteen päivitystavasta.

Prioriteetti: 3

K1.5 Metadatan siirto

Käyttäjä voi siirtää metadataa ohjelman eri instanssien välillä.

Prioriteetti: 3

K1.5.1 Raakadatan siirto

Käyttäjä voi siirtää alkuperäisessä muodossa olevaa raakadataa ohjelman eri instanssien välillä ja muissa sovelluksissa käsittelyä varten.

Prioriteetti: 3

K1.5.2 Atomilauseiden siirto

Käyttäjä voi siirtää raakadataa atomilausemuodossa ohjelman eri instanssien välillä.

Prioriteetti: ei toteuteta

K1.5.3 Siirrettävän metadatan valinta

Käyttäjä voi valita siirrettävät datat lähteiden perusteella.

Prioriteetti: 3

K2.1 Resurssiverkon integrointi

Käyttäjä voi tuottaa atomilauseista resurssiverkon käynnistämällä integraatioprosessin (resurssiverkon määritelmä liitteessä 1, integraation määritelmä liittessä 2).

Prioriteetti: 1

K2.2 Integroinnin mukauttaminen

Käyttäjä voi vaikuttaa integraatioprosessiin antamalla sille ehtoja.
Prioriteetti: 3

K2.2.1 Atomilauseiden valinta

Käyttäjä voi valita lähteen, josta saadut atomilauseet valitaan integraatioon. Lähteen koko, hakupäivämäärä, formaatti ja onko sen tietueita integroitu on käyttäjälle näkyvissä. Jos lähteen tietueita on jo integroitu, uusi integrointikerta integroi vain puuttuvat tietueet.
Prioriteetti: 3

K2.2.2 Vastaavuusarvojen asettaminen resursseille

Käyttäjä voi asettaa integraatioon kullekin resurssille oman prosenttimuotoisen vastaavuusarvon, jonka mukaan integraatio määrittää miten tarkkaan resurssien pitää vasta toisiaan.
Prioriteetti: ei toteuteta

K2.2.3 Integraation arvioitu kesto

Käyttäjä saa arvion integraation kestosta valittuaan siihen tulevat resurssit.
Prioriteetti: ei toteuteta

K2.3 Integraation toistettavuus

Käyttäjä voi ajaa integraation useaan kertaan ja muodostaa jokaisesta integroinnista uuden resurssiverkon tai lisätä resursseja olemassaolevaan verkkoon.
Prioriteetti: 1

K2.3.1 Useat resurssiverkot

Käyttäjällä voi olla järjestelmässä useita toisistaan riippumattomia resurssiverkkoja.
Prioriteetti: 3

K2.4 Integraation lokitiedot

Käyttäjä saa lokitiedot integraatiosta. Lokitietoihin kuuluu tieto siitä kuinka monta resurssia luotiin missäkin vaiheessa ja kuinka monta raakadatatietuetta käsiteltiin

missäkin vaiheessa. Lokitiedoissa kerrotaan myös tapaukset joissa integraatioalgoritmi (ks. liite 2) ei osannut tunnistaa olisiko tietueet pitänyt yhdistää samaksi resurssiksi vai ei.
Prioriteetti: ei toteuteta

K2.4.1 Lokitietojen siirto

Lokitiedot voi tallentaa tai ainakin siirtää leikepöydälle.
Prioriteetti: ei toteuteta

K2.5 Integraatiossa käytettävät minimiheuristiikat

Integraatiovaiheessa resurssien ja yhteyksien tunnistamisessa käytetään heuristiikkoja, jotka on määritelty integraatioalgoritmin kuvauksen yhteydessä (ks. Liite 2).
Prioriteetti: 1

K3.1 Resurssien hakeminen

Käyttäjä voi hakea resursseja hakuikkunan avulla.
Prioriteetti: 3

K3.1.1 Resurssien hakeminen ominaisuuksien perusteella

Käyttäjä voi hakea resursseja niiden ominaisuuksien perusteella (resurssien ominaisuudet on listattu resurssiverkon määrittelyssä liittessä 1).
Prioriteetti: 3

K3.1.2 Resurssien hakeminen yhteyksien perusteella

Käyttäjä voi hakea resursseja niiden yhteyksien lukumäärän perusteella (mahdolliset yhteydet on eri resursseille määritelty resurssiverkon määrittelyssä liitteessä 1). Tuettavat operaatiot ovat "yhteyksien lukumäärä > x" tai "yhteyksien lukumäärä < x". Käyttäjä voi myös hakea ylipäänsä ne resurssit joihin on jonkilaisia yhteyksiä.
Prioriteetti: 2

K3.1.3 Hakuehtojen muokkaus käsin

Käyttäjä voi muokata käyttöliittymän avulla syntyneitä SQL lausekkeen

hakuehtoja käsin ennen haun suorittamista. Hakuehtojen muuttaminen käyttöliittymästä nollaa käsintehdyt muokkaukset. Koko SQL lause näytetään, mutta käyttäjä voi muuttaa vain hakuehtoja.
Prioriteetti: 4

K3.2 Resurssien selaus

Käyttäjä voi saada listan haun tuloksina olevista resursseista ominaisuuksineen uuteen selausikkunaan. Listassa näytetään seuraavat ominaisuudet: kaikki muut tiedot paitsi (1) tieto mistä atomilausetietueista resurssi on muodostettu, (2) mitä ulkoisia tunnisteita resurssin liittyy ja (3) mitä hakusanoja resurssiin liittyy.
Prioriteetti: 2

K3.2.1 Hakutuloksen järjestäminen

Käyttäjä voi järjestää resurssien listan nousevaan tai laskevaan järjestykseen valitsemiensa ominaisuuksien mukaan (numeerisille suuruusjärjestys, kirjainjonoille aakkosjärjestys, päivämäärille aikajärjestys).
Prioriteetti: ei toteuteta

K3.2.2 Resurssien lisäys yhteysverkkonäkymään

Käyttäjä voi lisätä resursseja selausikkunasta yhteysverkkonäkymään.
Prioriteetti: ei toteuteta

K3.2.3 Hakutuloksen tallennus

Käyttäjä voi tallentaa hakutuloksen resurssien ominaisuuslistan formaattiin joka on aukaistavissa Microsoft Excel 2003 ja OpenOffice.org Calc 1.1.0 -ohjelmistoilla.
Prioriteetti: 2

K3.2.3.1 Valittujen ominaisuuksien tallennus hakutuloksesta

Käyttäjä voi valita mitkä ominaisuudet hakutuloksesta tallennetaan muotoon joka on aukaistavissa Microsoft Excel 2003 ja OpenOffice.org Calc 1.1.0 -ohjelmistoilla.
Prioriteetti: ei toteuteta

K3.2.4 Resurssin avaus selausikkunasta

Käyttäjä voi avata resurssin listasta resurssin ominaisuusikkunaan (resurssien ominaisuudet ja yhteydet on listattu resurssiverkon määrittelyssä luku 6.1).
Prioriteetti: ei toteuteta

K3.2.5 Hakutuloksen avaaminen selausikkunaan

Käyttäjä voi avata osan hakutuloksesta uuteen selausikkunaan.
Prioriteetti: ei toteuteta

K3.2.6 Hakutuloksen laajentaminen

Käyttäjä voi laajentaa hakutulosta niin, että listasta valitun resurssin valitun tyyppisen yhteyden päässä olevat resurssit otetaan hakutulokseen mukaan. Tämän jälkeen yhteystyypin kohdalla laajennetun resurssin väri vaihtuu, niin että käyttäjä näkee mitkä resurssin yhteystyypit hän on laajentanut.
Prioriteetti: ei toteuteta

K3.2.7 Resurssiverkon puhdistus

Käyttäjä voi poistaa resurssiverkosta tarkastelua häiritseviä resursseja. Käyttäjä määrittää poistettavat resurssit järjestelmän hakuominaisuuksien avulla. Käyttäjä voi määrittää hakuikkunaan ehdot poistettaville resursseille (esim. viittaukset resurssiin < 2). Käyttäjä voi valita koko hakutuloksen tai tämän osan poistettavaksi resurssi-ikkunassa. Kun resurssi poistetaan kaikki resurssiin liittyvät yhteydet poistetaan.
Prioriteetti: 3

K3.3 Resurssin tarkastelu

Käyttäjä voi nähdä yhden resurssin kaikki ominaisuudet ja yhteydet muunnettavissa kentissä ominaisuusikkunassa (resurssien ominaisuudet ja yhteydet on listattu resurssiverkon määrittelyssä liittessä 1)
Prioriteetti: ei toteuteta

K3.3.1 Resurssien ominaisuuksien ja yhteyksien muuttaminen

Käyttäjä voi muuttaa ja lisätä resurssien kaikkia ominaisuuksia ja yhteyksiä
Prioriteetti: ei toteuteta

K3.3.2 Uusien resurssien luonti

Käyttäjä voi luoda uusia resursseja
Prioriteetti: ei toteuteta

K3.4 Yhteysverkon tarkastelu

Käyttäjä voi visualisoida resurssien välisen yhteysverkon yhteysverkkoikkunaan. Käyttäjä voi valita kaikki tai osan hakutuloksen resursseista tarkasteluun.
Prioriteetti: ei toteuteta

K3.4.1 Yhteysverkon visuaalisuus

Jokaisesta verkon resurssista on nähtävissä tunniste ja resurssin typpi. Jokaisesta yhteydestä on nähtävissä yhteyden suunta. Resursseja tulee olla nähtävissä ainakin 300 kerrallaan.
Prioriteetti: ei toteuteta

K3.4.1.1 Yhteysverkon solmujen siirto

Käyttäjä voi siirtää solmujen paikkaa yhteysverkkoikkunassa.
Prioriteetti: ei toteuteta

K3.4.1.2 Tunnistetietojen säätäminen

Käyttäjä voi säätää resurssien tunnisteet näytettäviksi tai otettaviksi pois näytöstä.
Prioriteetti: ei toteuteta

K3.4.2 Resurssin avaus yhteysverkosta

Käyttäjä voi valita resurssien yhteysverkosta resurssin ominaisuusikkunaan
Prioriteetti: ei toteuteta

K3.4.3 Yhteysverkon tallennus Pajek-muodossa

Käyttäjä voi tallentaa yhteysverkon Pajek -listaformaattiin. Käyttäjä valitsee

tallennettavan (yhden) yhteystyypin.
Prioriteetti: 1

K3.4.4 Yhteysverkon tallennus matriisina

Käyttäjä voi tallentaa yhteysverkon formaattiin joka ilmaisee yhteysverkon matriisina ja on aukaistavissa aukaistavissa Microsoft Excel 2003 ja OpenOffice.org Calc 1.1.0 -ohjelmistoilla. Käyttäjä voi valita tallennettavan (yhden) yhteystyypin.
Prioriteetti: ei toteuteta

K3.5 Merkitysverkon luonti

Käyttäjä voi luoda yhteysverkon päälle resurssien välisen merkitysverkon. Merkitysverkko koostuu resursseille annetuista painoarvoista ja niiden välisistä merkitysyhteyksistä joilla on painoarvo.
Prioriteetti: ei toteuteta

K3.5.1 Merkitysverkon merkitysten asetus

Käyttäjä voi asettaa resursseille nimettyjä merkityksiä asettamalla resurssille painoja selausikkunassa ja yhteysverkossa. Painot ilmaistaan positiivisella tai negatiivisella reaaliluvulla.
Prioriteetti: ei toteuteta

K3.5.1.1 Resursseihin liittyvien merkitysten näyttö

Selausikkunassa näytetään resursseihin mahdollisesti liittyvät merkitykset.
Prioriteetti: ei toteuteta

K3.5.2 Merkitysverkon yhteyksien luonti

Käyttäjä voi luoda resurssien välille suunnatun ja nimetyn merkitysyhteyden sekä asettaa sille painon positiivisella tai negatiivisella reaaliluvulla. Kahden resurssin välillä voi olla useita eri tyyppisiä merkitysyhteyksiä.
Prioriteetti: ei toteuteta

K3.5.2.1 Merkitysverkon merkitysten poisto

Käyttäjä voi poistaa resursseille asetettuja merkityksiä sekä merkitysverkon yhteyksiä.
Prioriteetti: ei toteuteta

K3.5.3 Useat merkitysverkot

Ohjelmalla voi käsitellä useita yhteen resurssiverkkoon liittyviä useita merkitysverkkoja.
Prioriteetti: ei toteuteta

K3.5.3.1 Käsiteltävänä yksi merkitysverkko kerrallaan

Ohjelma pystyy näyttämään yhden merkitysverkon kerrallaan.
Prioriteetti: ei toteuteta

K3.5.4 Merkitysverkkojen tallennus

Käyttäjä voi tallentaa merkitysverkkoja
Prioriteetti: ei toteuteta

K3.5.4.1 Merkitysverkkojen tallennus Pajek-muodossa

Käyttäjä voi tallentaa merkitysverkkoja Pajek -listaformaattiin
Prioriteetti: ei toteuteta

K3.5.4.2 Merkitysverkkojen tallennus matriisina

Käyttäjä voi tallentaa merkitysverkkoja formaatissa joka ilmaisee merkistysverkon matriisina ja on aukaistavissa Microsoft Excel 2003 ja OpenOffice.org Spreadsheet 1.1.0 -ohjelmistoilla.
Prioriteetti: ei toteuteta

K3.5.5 Merkitysverkkojen lataus

Käyttäjä voi ladata tallennetun merkitysverkon Pajek-listaformaattia olevista tiedostoista. Merkitysverkkoa ladattaessa olemassaolevaan verkkoon liitetään vain lisättävässä verkossa olevat uudet merkitykset.
Prioriteetti: ei toteuteta

K3.5.5.1 Yhtesverkon resurssien lataus ja merkitysten karsinta

Merkitysverkon latauksen yhteydessä käyttäjä voi valita ladataanko yhteysverkoon ne puuttuvat resurssit joihin viitataan merkitysverkkosta vai poistetaanko merkitysverkosta ne merkitykset ja yhteydet joihin ei löydy yhteysverkon resurssia.
Prioriteetti: ei toteuteta

U1 Tallennettavan matriisin suunnan valinta

Käyttäjä voi valita tallennetaanko vaatimuksen kuvaileman K3.4.4 matriisin subjekti ja objekti -parit riveittäin vai sarakkeittain.
Prioriteetti: ei toteuteta

U2 Resurssien poisto selauksesta

Käyttäjä voi valita hakutuloksesta resursseja jotka poistetaan selaustuloksesta
Prioriteetti: 1

U3 Visualisoinnin automatisointi

Jos käyttäjä päivittää laajennusta (K3.2.6), yhteysverkon visualisointi (K3.4.1) päivittyy automaattisesti.
Prioriteetti: ei toteuteta

U4 Tietyn tyyppisten yhteyksien tallennus

Käyttäjä voi valita että hakutuloksesta tallennetaan tiedostoon (K3.2.3, K3.4.3) tietyn tyyppiset tai kaikki yhteydet
Prioriteetti: 1

3.2. Suorituskykyvaatimukset

S1 Resurssiverkon koko

Resurssiverkon maksimisolmumäärä on 2147483647 (Javan suurin int arvo).

S2 Selauksen nopeus
Selausikkunan vieritys vierityskäskystä vierityksen valmistumiseen saa kestää maksimissaan 0,1 sek.

3.3. Laadulliset vaatimukset

L1 Ympäristö
Järjestelmä toimii laitoksen Linux-ympäristössä.

L2 Järjestelmän ulkoiset komponentit
Kaikkien järjestelmän käyttämien ulkoisten komponenttien on oltava vapaan ohjelmistolisenssin alaisia (open source) tai muuten maksuttomia.

L3 Järjestelmän asennus
Asiakkaan tulee pystyä asentamaan järjestelmän laitoksen Linux-koneelle joka ei ole yhteydessä verkkoon.

L4 Järjestelmän iteraatio
Ohjelmiston integraatiovaiheen vaatimukset muuttuvat kun ensimmäinen versio ohjelmistosta on ollut asiakkaan arvioitavissa. Tällöin ohjelmiston vaatimuksille tehdään toinen iteraatio. Iteraatiossa oletetaan erityisesti että integrointiin liittyvät menetelmät tulevat muuttumaan integroinnissa käytettävien algoritmien osalta. 1. Toteutus tehdään viimeistään 21.11. mennessä. Asiakas antaa muutosehdotuksen kolmen arkipäivän sisällä 1. toteutuksen valmistumisesta. Muutosten toteuttamiseen varataan 50 työtuntia.
Huomautus: tätä vaatimusta ei toteuteta aikataulun tiukkuuden takia.

3.4. Vaatimusten prioriteetit

Prioriteetti 1
K1.1 Valmiin raakadatan tuonti
K1.1.5 oai_citeseer-muotoisten tiedostojen tuonti

K1.1.6 XML Dublin Core-muotoisten tiedostojen tuonti

K1.2 Itse tehdyn metadatan tuonti

K1.2.1 Pikaformaatti-tiedostojen tuonti

K2.1 Resurssiverkon integrointi

K2.3 Integraation toistettavuus

K2.5 Integraatiossa käytettävät minimiheuristiikat

K3.2.5 Hakutuloksen avaaminen selausikkunaan

K3.4.3 Yhteysverkon tallennus Pajek-muodossa

U2 Resurssien poisto selauksesta

U4 Tietyn tyyppisten yhteyksien tallennus

Prioriteetti 2

K1.1.1 XML Dublin Core harvestointi

K1.3 Metadatan säilytys

K1.3.1 Raakadatan säilytys

K3.1.2 Resurssien hakeminen yhteyksien perusteella

K3.2 Resurssien selaus

K3.2.3 Hakutuloksen tallennus

Prioriteetti 3

K1.3.2 Atomilauseiden säilytys

K1.4 Lähteiden päivitys

K1.5 Metadatan siirto

K1.5.1 Raakadatan siirto

K1.5.3 Siirrettävän metadatan valinta

K2.2 Integroinnin mukauttaminen

K2.2.1 Atomilauseiden valinta

K2.3.1 Useat resurssiverkot

K3.1 Resurssien hakeminen

K3.1.1 Resurssien hakeminen ominaisuuksien perusteella

K3.2.7 Resurssiverkon puhdistus

Prioriteetti 4

K3.1.3 Hakuehtojen muokkaus käsin

Ei toteuteta

K1.1.2 Pajek-tiedostojen tuonti

K1.1.3 DBLP-tiedostojen tuonti

K1.1.4 BibTeX metadatan tuonti

K1.1.4.1 BibTeX tiedostojen tuonti

K1.1.4.2 BibTeX tietokantaluku

K1.5.2 Atomilauseiden siirto

K2.2.2 Vastaavuusarvojen asettaminen resursseille

K2.2.3 Integraation arvioitu kesto

K2.4 Integraation lokitiedot

K2.4.1 Lokitietojen siirto

K3.2.1 Hakutuloksen järjestäminen

K3.2.2 Resurssien lisäys yhteysverkkonäkymään

K3.2.3.1 Valittujen ominaisuuksien tallennus hakutuloksesta

K3.2.4 Resurssin avaus selausikkunasta

K3.2.6 Hakutuloksen laajentaminen

K3.3 Resurssin tarkastelu

K3.3.1 Resurssien ominaisuuksien ja yhteyksien muuttaminen

K3.3.2 Uusien resurssien luonti

K3.4 Yhteysverkon tarkastelu

K3.4.1 Yhteysverkon visuaalisuus

K3.4.1.1 Yhteysverkon solmujen siirto

K3.4.1.2 Tunnistetietojen säätäminen

K3.4.2 Resurssin avaus yhteysverkosta

K3.4.4 Yhteysverkon tallennus matriisina

K3.5 Merkitysverkon luonti

K3.5.1 Merkitysverkon merkitysten asetus

K3.5.1.1 Resursseihin liittyvien merkitysten näyttö

K3.5.2 Merkitysverkon yhteyksien luonti

K3.5.2.1 Merkitysverkon merkitysten poisto

K3.5.3 Useat merkitysverkot

K3.5.3.1 Käsiteltävänä yksi merkitysverkko kerrallaan

K3.5.4 Merkitysverkkojen tallennus

K3.5.4.1 Merkitysverkkojen tallennus Pajek-muodossa

K3.5.4.2 Merkitysverkkojen tallennus matriisina

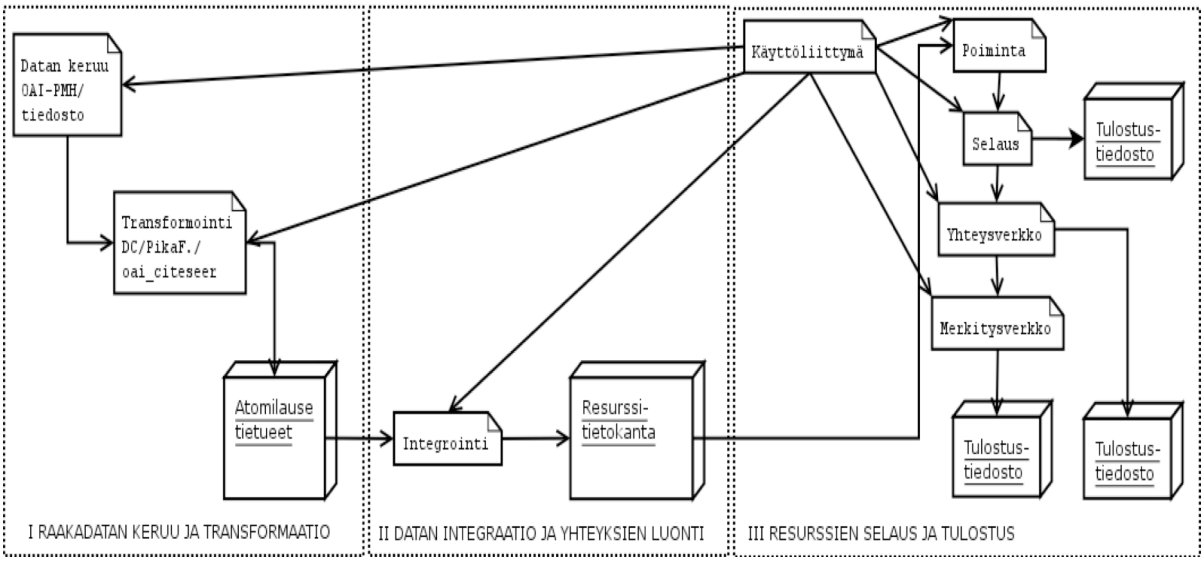
K3.5.5 Merkitysverkkojen lataus

K3.5.5.1 Yhteysverkon resurssien lataus ja merkitysten karsinta

U1 Tallenettavan matriisin suunnan valinta

U3 Visualisoinnin automatisointi

4. Järjestelmäarkkitehtuuri



Kuva 1. Arkkitehtuurikaavio

Järjestelmä koostuu kolmesta osajärjestelmästä.

Raakadatan keruu ja transformointi

Järjestelmään kuuluu datan keruu- ja transformointikomponentti, jonka avulla järjestelmään voidaan tuoda yllä määriteltyjä kolmea eri metadataformaattia. Transformointikomponenteista kukin transformoi yhden tyyppistä metadataa yhteiseen muotoon ja tallentaa syntynyttä atomilausedataa tiedostoon.

Integrointi

Osajärjestelmä lukee transformointi-vaiheen tuottamaa atomilausedataa ja luo atomilauseista resurssiverkon, joka tallennetaan tietokantaan.

Selaus ja poiminta

Osajärjestelmä koostuu hakukomponentista, jonka avulla toteutetaan poimintojen tekeminen integrointi-vaiheessa syntyneestä resurssiverkosta. Selauskomponentti mahdollista poiminnan tuloksen tarkastelun listana ja yksittäisen resurssin ominaisuuksia tarkastelun ja tietojen muokkaamisen. Lista voidaan tulostaa tiedostoon. Yhteysverkkokomponentti esittää poiminnan tuloksen graafisessa muodossa. Poiminnan tulos voidaan tallentaa tiedostoon PAJEK-muodossa tai yhteysmatriisina tekstitiedostoon.

5. Raakadatan Transformointi

5.1. Transformoitavien atomilauseiden arvojen kanonisointi

- Jos datankeruukomponentti ei tiedä käytettävää merkistöä, oletetaan sen olevan UTF-8
- Jos merkkijonossa on :// tai www., niin kanonisointia ei suoriteta ja tyypiksi tulee joku
- Poistetaan merkkijonon alusta ja lopusta kaikki whitespace (Javan Character.isWhitespace() mukaan)
- Muutetaan kaikki peräkkäiset whitespace-merkit yhdeksi välilyönniksi
- Poistetaan kaikki merkit paitsi a-z A-Z 0-9 , . / - : ~
- Muutetaan kaikki a-z -> A-Z
- Henkilöiden ja organisaatioiden käsittely:
- Jos merkkijonossa on " INC", " LTD", " OF ", " CO.", " CORP", " PRESS ", " UNIV", " PUBL" kyseessä on organisaatio
- Jos merkkijonossa on " ED.", " EDS.", "ET AL.", "ET ALII.", " ANON., " JR.", " SR." poistetaan ne ja kyseessä on henkilö(itä)
- Jos merkkijonossa on " X.", missä X on kirjain, kyseessä on henkilö(itä)
- Jos ei ollut organisaatio eikä henkilö, tyyppi joku
- Henkilönimet lisäkanonisoidaan seuraavasti:
    - Muutetaan muotoon SUKUNIMI(+,+välilyönti+ensimmäisen etunimen ensimmäinen kirjain)
    - Jos kyseessä on oai_citebase-muotoista dataa otetaan merkkijonon viimeisen välilyönnin jälkeinen merkkijono, lisätään ", " ja merkkijonon ensimmäinen kirjain.
    - Muulloin merkkijonosta poistetaan ", X"-merkkijonon (missä X on kirjain) jälkeen tulevat merkit. Jos merkkijonossa ei ole pilkkua, lisätään ensimmäisen välilyönnin eteen pilkku ja poistetaan tämän välilyönnin ja sitä seuraavan merkin jälkeen tulevat merkit

        Esim. 1: LI

        Esim. 2: GATES, W

        Esim. 3: VAN DER SAAR, E

        Esim. 4: ALA-MKINEN, E

        Esim. 5: CHING A, J
- Organisaatiot ja jotkut tulevat sellaisenaan

Esim. 1: MICROSOFT

Esim. 2: MICROSOFT PRESS

Esim. 3: MICROSOFT, INC.

5.2 Transformoidun datan tallennus

Dataa käsitellään ja tallennetaan järjestelmässä Unicode (UTF-8) -merkistöllä. Tallennettu data sisältää seuraavat tiedot:

Kuva 2. Tallennetun datan hierarkia

5.3. Yhteinen sanasto

Raakadatan transformoinnissa jokaisen datalähteen sanasto muutetaan yhteen yhteiseen sanastoon, jotta eri lähteiden vastaavat tietueet voidaan muuttaa yksikäsitteisiksi. Kun järjestelmään lisätään uusi tietolähde, sanastosta otetaan käyttöön soveltuvat termit. Jos uudessa tietolähteessä on termejä, joita vastaavaa termiä ei sanastosta löydy lisätään sanastoon soveltuva yleisluontoinen termi ja otetaan tämä käyttöön myös seuraavissa tietolähteissä.

Käytettävä sanasto:

Nimeke, KanoNimeke, TekijäNimi, HenkilöTekijä, OrganisaatioTekijä, JokuTekijä, Hakusanat, JulkaisijaNimi, JulkaisuNimi, KanoJulkaisuNimi, HenkilöJulkaisija, OrganisaatioJulkaisija, JokuJulkaisija, AvustajaNimi, HenkilöAvustaja, OrganisaatioAvustaja, JokuAvustaja, Julkaisupäivämäärä, Formaatti, URLTunniste, URNTunniste, OAITunniste, OpenURLTunniste, DOITunniste, INFOTunniste, URITunniste, SisäinenTunniste Viittaustunniste, Kieli, Yhteys, Viitattu, Viittaava, Oikeuksienomistaja, HenkilöOikeuksienomistaja, OrganisaatioOikeuksienomistaja, JokuOikeuksienomistaja, JokuToimija, HenkilöToimija, OrganisaatioToimija, JokuRooli, JokuKanava, LehtiKanava, KonferenssiKanava, JulkaisusarjaKanava, RaporttisarjaKanava, TietokantaKanava, Aihe, Hallinnoija

5.4. Transformaatiosäännöt

Transformaatio: Dublin Core Simple

Title

OminaisuudenNimi: Nimeke

OminaisuudenArvo: Merkkijono sellaisenaan

OminaisuudenNimi: KanoNimeke

OminaisuudenArvo: Kanonisoitu merkkijono, josta lisäksi poistettu alle 4-kirjaimiset välilyönnillä erotellut merkkijonot. Maksimikoko: 255 merkkiä.

Creator

OminaisuudenNimi: HenkilöTekijä/OrganisaatioTekijä/JokuTekijä

OminaisuudenArvo: Merkkijono sellaisenaan

Subject

OminaisuudenNimi: Aihe

OminaisuudenArvo: Kanonisoitu merkkijono

Description

OminaisuudenNimi: Hakusanat

OminaisuudenArvo: Kanonisoitu merkkijono

Publisher

OminaisuudenNimi: HenkilöJulkaisija/OrganisaatioJulkaisija/JokuJulkaisija

OminaisuudenArvo: Merkkijono sellaisenaan

Contributor

OminaisuudenNimi: HenkilöAvustaja/OrganisaatioAvustaja/JokuAvustaja

OminaisuudenArvo: Merkkijono sellaisenaan

Date

OminaisuudenNimi: Julkaisupäivämäärä

OminaisuudenArvo: Jos issued-tarkennetta ei ole (kts, alla), niin arvo sellaisenaan (muotoa YYYY-MM-DD)

Type

OminaisuudenNimi: Hakusanat

OminaisuudenArvo: Kanonisoitu merkkijono

Format

OminaisuudenNimi: Formaatti

OminaisuudenArvo: Kanonisoitu merkkijono

Identifier

OminaisuudenNimi:

Jos (kanonisoimaton merkkijonoarvo isoilla tai pienillä) alkaa http, niin OminaisuudenNimi URLTunniste

Jos alkaa urn, niin OminaisuudenNimi URNTunniste

Jos alkaa info:oai, niin OminaisuudenNimi OAITunniste

Jos alkaa info:ofi, niin OminaisuudenNimi OpenURLTunniste

Jos alkaa info:doi, niin OminaisuudenNimi DOITunniste

Jos alkaa info (eikä jatku edellisillä), niin OminaisuudenNimi INFOTunniste

Jos sisältää vuosiluvun (neljä peräkkäistä numeroa väliltä 1800-2010) alussa tai lopussa, yli 2 pistettä tai pilkkua ja yli 16 merkkiä Viittaustunniste

Jos ei em. ja sisältää ://, niin URITunniste

muuten SisäinenTunniste

OminaisuudenArvo Merkkijono sellaisenaan

Source

Ohitetaan

Language

OminaisuudenNimi: Kieli

OminaisuudenArvo: ISO 639 Merkkijono kanonisoituna ilman viivalla erotettua lisätunnistetta (kts. http://www.w3.org/WAI/ER/IG/ert/iso639.htm)

Relation

Ominaisuudennimi: Yhteys

Ominaisuudenarvo: Merkkijono sellaisenaan

Coverage

Ohitetaan

Rights

Ohitetaan

Transformaatio: oai_citeseer

Transformoinnissa toteutetaan kaikki Dublin Core Simple elementit sekä seuraavat oai_citeseer -elementit:

identifier

OminaisuudenNimi: SisäinenTunniste

OminaisuudenArvo: Merkkijono sellaisenaan

oai_citeseer:author > name

OminaisuudenNimi: HenkilöTekijä

OminaisuudenArvo: Merkkijono sellaisenaan

oai_citeseer:author > affiliation

OminaisuudenNimi: OrganisaatioTekijä

OminaisuudenArvo: Kanonisoitu merkkijono

oai_citeseer:relation type="References"

OminaisuudenNimi: Viitattu

OminaisuudenArvo: Merkkijono sellaisenaan

oai_citeseer:relation type="Is Referenced By"

OminaisuudenNimi: Viittaava

OminaisuudenArvo: Merkkijono sellaisenaan

Transformaatio: Dublin Core Qualified

Transformoinnissa toteutetaan kaikki Dublin Core Simple elementit sekä seuraavat DCTERMS -elementit:

dcterms:bibliographicCitation

OminaisuudenNimi: Viittaustunniste

OminaisuudenArvo: Merkkijono kanonisoituna

dcterms:issued

OminaisuudenNimi: Julkaisupäivämäärä

OminaisuudenArvo: Arvo muutettuna YYYY-MM-DD muotoon

dcterms:references

OminaisuudenNimi: Viitattu

OminaisuudenArvo: Merkkijono sellaisenaan

dcterms:is referenced by

OminaisuudenNimi: Viittaava

OminaisuudenArvo: Merkkijono sellaisenaan

dcterms:rightsHolder

OminaisuudenNimi: HenkilöOikeuksienomistaja/OrganisaatioOikeuksienomistaja/JokuOikeuksienomistaja

OminaisuudenArvo: Merkkijono sellaisenaan

Transformaatio: Pikaformaatti dokumenteille

Tietueen Tunnisterivi

OminaisuudenNimi: SisäinenTunniste

OminaisuudenArvo: Merkkijono sellaisenaan

1 - Tekijät

OminaisuudenNimi: HenkilöTekijä/OrganisaatioTekijä/JokuTekijä

OminaisuudenArvo: Merkkijono sellaisenaan

2 - Nimeke

OminaisuudenNimi: Nimeke

OminaisuudenArvo: Merkkijono sellaisenaan

3 - Julkaisija

OminaisuudenNimi: HenkilöJulkaisija/OrganisaatioJulkaisija/JokuJulkaisija

OminaisuudenArvo: Merkkijono sellaisenaan

4 - Julkaisu

OminaisuudenNimi: JulkaisuNimi

OminaisuudenArvo: Merkkijono sellaisenaan

5 - JulkaisuVuosi

OminaisuudenNimi: Julkaisupäivämäärä

OminaisuudenArvo: Päivämäärä muutettuna YYYY-MM-DD -muotoon

6 - ViittausTunniste

OminaisuudenNimi: Viittaustunniste

OminaisuudenArvo: Merkkijono kanonisoituna

7 - Viitatut

OminaisuudenNimi: Viitattu

OminaisuudenArvo: Merkkijono sellaisenaan

8 - Viittaavat

OminaisuudenNimi: Viittaava

OminaisuudenArvo: Merkkijono sellaisenaan

9 - Hakusanat

OminaisuudenNimi: Hakusanat

OminaisuudenArvo: Kanonisoitu merkkijono

Transformaatio: Pikaformaatti nimille

2 - Toimija, tyyppi tuntematon
OminaisuudenNimi: TuntematonToimija
OminaisuudenArvo: Merkkijono sellaisenaan

2.1 - Henkilö
OminaisuudenNimi: HenkilöToimija
OminaisuudenArvo: Merkkijono sellaisenaan

2.2 - Organisaatio
OminaisuudenNimi: ToimijaNimi
OminaisuudenArvo: Merkkijono sellaisenaan
OminaisuudenNimi: OrganisaatioToimija
OminaisuudenArvo: Merkkijono kanonisoituna

3 - Rooli, tyyppi tuntematon
OminaisuudenNimi: JokuRooli
OminaisuudenArvo: Merkkijono sellaisenaan

3.1 - Julkaisija
OminaisuudenNimi: JokuJulkaisija
OminaisuudenArvo: Merkkijono sellaisenaan

4 - Kanava, tyyppi tuntematon
OminaisuudenNimi: JokuKanava
OminaisuudenArvo: Merkkijono sellaisenaan

4.1 - Lehti
OminaisuudenNimi: LehtiKanava
OminaisuudenArvo: Merkkijono sellaisenaan

4.2 - Konferenssi
OminaisuudenNimi: KonferenssiKanava
OminaisuudenArvo: Merkkijono sellaisenaan

4.3 - Julkaisusarja

OminaisuudenNimi: JulkaisusarjaKanava

OminaisuudenArvo: Merkkijono sellaisenaan

4.4 - Raporttisarja

OminaisuudenNimi: RaporttisarjaKanava

OminaisuudenArvo: Merkkijono sellaisenaan

4.5 - Tietokanta

OminaisuudenNimi: TietokantaKanava

OminaisuudenArvo: Merkkijono sellaisenaan

Arvon jälkeinen kaksoispiste

OminaisuudenNimi: OrganisaatioHallinnoija

OminaisuudenArvo: Merkkijono sellaisenaan

Arvon jälkeisen pystyviivan | jälkeen pilkulla erotellut merkkijonot

OminaisuudenNimi: Aihe

OminaisuudenArvo: Merkkijono kanonisoituna

¤ -merkillä eroteltu merkkijono

OminaisuudenNimi: SisäinenTunniste

OminaisuudenArvo: Merkkijono kanonisoituna

6. Integraatiovaihe

6.1. Resurssiverkon määrittely

Resurssiverkko koostuu resursseista ja niiden välisistä yhteyksistä. Resursseja on neljää tyyppiä: toimijat, kanavat, roolit ja dokumentit.

- Kunkin toimijan ja toisen resurssin väliseen yhteyteen liittyy jokin rooli.

- Jokaiseen resurssiin ja resurssien väliseen yhteyteen voi liittää painotettuja merkityksiä.
- Jokaiselle resurssille annetaan yksikäsitteinen järjestelmän sisäinen tunniste, joka muodostetaan resurssin nimestä ja juoksevasta numerosta. Jokaiseen resurssiin voi liittyä myös kolme eri hakusanaa.

Toimija

Toimijat ovat henkilöitä, organisaatioita tai muita toimijoita. Kuhunkin toimijaan voi liittyä useita kanavia ja dokumentteja. Toimijat liittyvät dokumentteihin jonkin roolin kautta.

Kanava

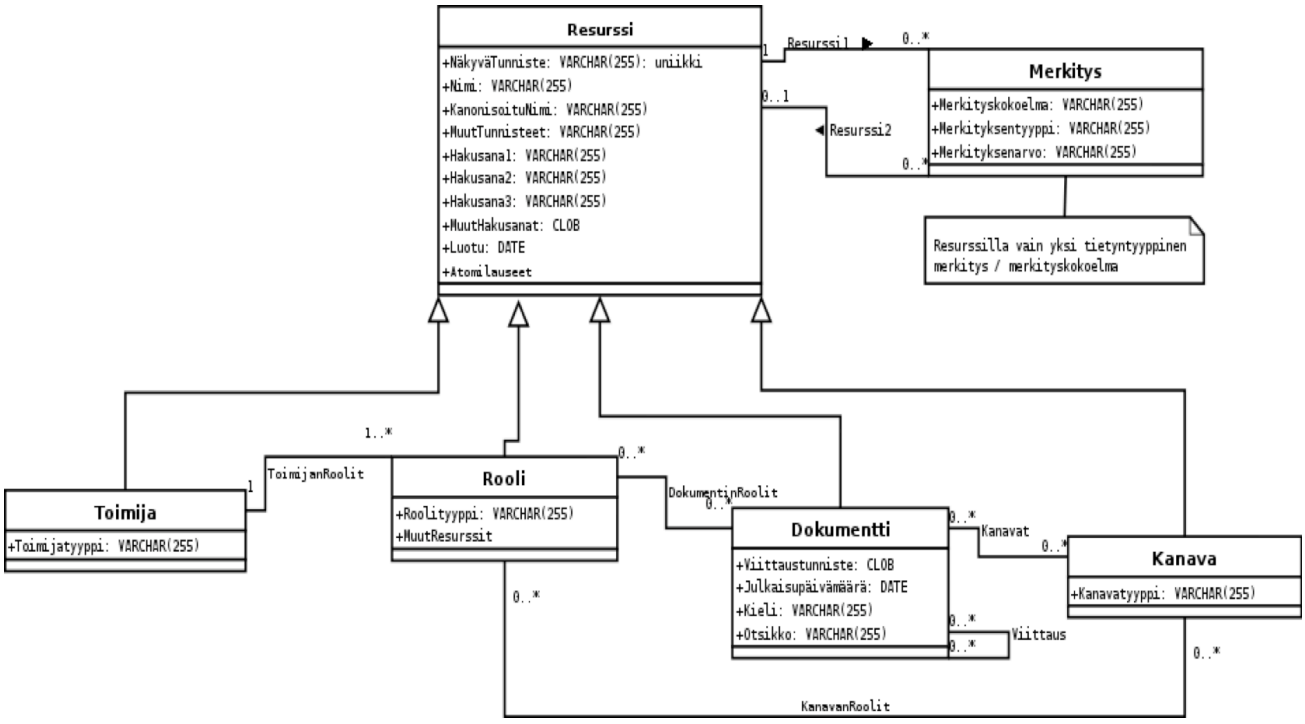
Kanavat ovat julkaisukanavia (esim. julkaisusarjoja). Jokaisella kanavalla on sijainti. Niistä on yhteys nollaan tai useampaan toimijaan, joka on kanavan hallinnoija. Kanavasta on myös yhteys nollaan tai useampaan dokumenttiin.

Dokumentti

Dokumenttiin liittyy tieto sen esitystyypistä, kielestä, formaatista sekä dokumentin kuvaus. Jokaisesta dokumentista on myös yhteys nollaan tai enempään toimijaan jonka rooli on tekijä, toimijaan jonka rooli on avustaja ja toimijaan jonka rooli on oikeudenomistaja. Dokumentista on myös yhteys nollaan tai useampaan kanavaan, viittaavaan dokumenttiin ja viitattuun dokumenttiin.

Merkitys

Merkitys voidaan liittää joko resurssien väliseen yhteyteen tai resursseihin. Merkityksen paino ilmaistaan reaaliluvulla. Merkitys liittyy aina johonkin merkitysverkkoon.



Kuva 3. UML-kaavio resurssiverkosta sekä siihen liittyvästä merkitysverkosta

6.2 Integraatioalgoritmin määrittely

Integraatiovaiheen syötteenä käytetään transformointivaiheessa luotuja atomilausetietueita.

Integraatioprosessi on jaettavissa seuraaviin vaiheisiin, jotka suoritetaan luetellussa järjestyksessä.

1. Dokumenttiresurssien luonti dokumenttitietueista

- Jokaisesta käsiteltävästä tietueesta luodaan dokumenttiresurssi tai viite jo olemassa olevaan vastaavaan dokumenttiresurssin.
- Dokumenttiresurssi sisältää käsiteltävän tietueen tiedot ja viitteen tähän tietueeseen.
- Jos tietuetta vastaava dokumenttiresurssi on jo olemassa, lisätään dokumenttiresurssiin viite tietueeseen sekä tietueen sisältämät uudet tiedot niin, että dokumenttiresurssissa jo olemassaolevia tietoja ei korvata.
- Dokumenttiresurssiin tallenetaan seuraavat tietueen tiedot: Nimeke > Nimi, KanoNimeke > KanonisoituNimi, Otsikko > Otsikko, URNTunniste, OAITunniste, OpenURLTunniste, DOITunniste, INFOTunniste, URITunniste, Viittaustunniste,

SisäinenTunniste > MuutTunnisteet, Julkaisupäivämäärä > Julkaisupäivämäärä, Kieli > Kieli, Aiheet (ensimmäiset kolme) > Hakusana 1 2 ja 3, Hakusanat > MuutHakusanat, Viitattu > Viitatut, Viittaava > Viittaavat

- Dokumenttien vastaavuus-heuristiikka: Jos tietueen KanoNimeke (verrataan yhtenäiseen muotoon muokattuja merkkijonoja) on sama kuin vertailtavan dokumenttiresurssin, niin dokumentit ovat samat. Tietueet ovat samat, jos vertailtavat merkkijonot ovat identtisiä.
- Jos luodaan uusi dokumenttiresurssi, tälle luodaan NäkyväTunniste joka on muotoa: Tekijän 3 ensimmäistä merkkiä + Julkaisuvuosi + ensimmäinen otsikon sana, joka > 3 merkkiä. Jos tunniste on jo käytössä, lisätään loppuun juokseva numerointi
- Tulos: jokaista tietuetta vastaa yksi dokumenttiresurssi ja yhteen dokumenttiresurssiin voi liittyä useita tietueita.

2. Muiden kuin dokumenttityyppisten resurssien luonti dokumenttitietueista

- Dokumenttiresursseista poimitaan tiedot uusien ei-dokumenttityyppisten resurssien luontiin. Uudet resurssit ja näiden vastaavuudet tunnistetaan heuristiikkojen avulla:
- Toimija-heuristiikka: Jokainen seuraavista tietueen atomilauseista käsitellään. Jos tietueessa oleva toimijan nimi on sama kuin resurssiverkossa oleva toimijan nimi, niin toimijat ovat samat. Tällöin liitetään resurssiin viite käsiteltävänä olevaan tietueeseen. Tämän lisäksi löydettyyn resurssiin liitetään yhteys käsiteltävää tietuetta vastaavaan dokumenttiresurssiin heuristiikan tunnistaman Roolin kautta.

  HenkilöTekijä, OrganisaatioTekijä, JokuTekijä,HenkilöJulkaisija, OrganisaatioJulkaisija, JokuJulkaisija, HenkilöAvustaja, OrganisaatioAvustaja, JokuAvustaja, HenkilöOikeuksienomistaja, OrganisaatioOikeuksienomistaja, JokuOikeuksienomistaja.

  Jos heuristiikka ei löydä vastaavaa reusurssia, luodaan uusi resurssi resurssiverkkoon.
- Kanava-heuristiikka: Heuristiikka käsittelee samoin kuin toimija-heuristiikka seuraavat tietueen atomilauseet: JokuKanava, LehtiKanava, KonferenssiKanava, JulkaisusarjaKanava, RaporttisarjaKanava, TietokantaKanava.
- Jos luodaan uusi resurssi, tälle luodaan NäkyväTunniste joka on muotoa: resurssin nimi sellaisenaan. Jos tunniste on jo käytössä, lisätään loppuun juokseva numerointi

3. Muiden kuin dokumenttityyppisten resurssien luonti ei-dokumenttitietueista

- Käsitellään tietueet jotka kuvaavat muun tyyppisiä resursseja kuin dokumentteja jotka on ladattu nimi-pikaformaatin avulla.

- Tietueista luodaan uusia resursseja jotka sisältävät tietueen tiedot ja viitteen tietueeseen.
- Heuristiikka: Heuristiikka vertailee kaikkia nimi-pikaformaatin tuottamia kanonisoituun muotoon muutettuja atomilauseita. Jos atomilauseessa kuvatun toimijan tai kanavan nimi on sama (verrataan yhtenäiseen muotoon muokattuja merkkijonoja) niin toimijat / kanavat ovat sama. Tällöin luodaan viite resurssista käsiteltävään tietueeseen
- Jos resurssi on jo olemassa heuristiikan mukaan, niin liitetään käsiteltävän tietueen tiedot resurssiin niin, että tietueessa olevat tiedot korvaavat resurssissa olevat tiedot. Luodaan viite aikaisemmin luotuun resurssiin.
- Jos luodaan uusi resurssi, tälle luodaan NäkyväTunniste joka on muotoa: resurssin nimi sellaisenaan. Jos tunniste on jo käytössä, lisätään loppuun juokseva numerointi

4. Dokumenttiresurssien välisten yhteyksien luominen

Dokumenttiresurssien väliset yhteydet tunnistetaan ja luodaan dokumenttiresurssien sisältämien viitetietojen (reference) perusteella käyttäen hyväksi dokumenttitietueisiin talletettuja käsiteltävän dokumentin alkuperäisiä viitetietoja

- Jokainen dokumenttiresurssi käsitellään nousevassa järjestyksessä Julkaisupäivämäärän mukaan.
- Viitattu-heuristiikka: Käsittele dokumenttiresurssiin liittyvät kaikki Viitattu-tyyppiset atomilauseet. Suorita jokaiselle atomilauseelle haku dokumenttiresurssin MuutTunnisteet kenttään. Jos dokumenttiresurssi löytyy, luo Viitattu-yhteys löydettyyn dokumenttiin. Jos ei löydy kirjoitetaan loki ja luodaan uusi dokumentti, jonka tietona on tämä Viitattu-tunniste sekä mahdollisesti Nimi ja KanoNimi jos nämä tiedot saatavilla Viittattu-kentästä. Luodaan yhteys tähän dokumenttiresurssiin.
- Viittaavat-heuristiikka: Vastaava, mutta luo käsiteltävään dokumenttiin Viittaava-yhteyden

5. Tilastot

- Tulostetaan tiedot integraation yhteydessä luoduista uusista resursseista järjestettynä KanonisoituNimi -ominaisuuden mukaan.
- Jokaisesta luodusta resurssista tulostetaan resurssin tunniste ja KanonisoituNimi. Jos luodusta resurssista ei ole yheyksiä muihin resursseihin, tämä ilmaistaan tulostuksen yhteydessä.
- Tulostetaan yleistä tilastotietoa: Luotujen resurssien lkm, käsiteltyjen raakadatatietueiden lkm.

7. Liitteet

Liite 1: Dokumenttityyppisen pikaformaatin määrittely

Pikaformaatti on yksinkertainen tapa esittää metadataa. Se on tarkoitettu käyttäjälle metadatan luomista varten. Pikaformaatin muoto on seuraavanlainen:

- Merkistönä UTF-8, rivinerotin LF-merkki
- #-merkillä alkava rivi on kommenttirivi
- Kaikki tyhjät ja kommenttirivit jätetään huomiotta
- Tiedoston ensimmäisellä rivillä on kaikkien seuraavien tietueiden tyyppi, joka voi olla

  1=Dokumentti
- Jos tiedosto on nimityyppistä pikaformaattia ensimmäisen rivin arvo on <> 1 (ks. Liite 2.)
- Toisella tiedoston rivillä on lähteen nimi
- Tietue alkaa rivillä, jolla kolme viivaa: "---"
- Tietueen toisella rivillä on tietueen tunniste
- Loput tietueen rivit ovat attribuuttirivejä
  - attribuuttirivit alkavat attribuutin numerolla, jonka jälkeen arvo
  - arvottomat attribuuttirivit ovat mahdollisia, mutta käsittelyssä ne ohitetaan
  - attribuuttirivit ovat attribuuttinumerojärjestyksessä
  - Jos attribuuttiarvo on moniarvoinen, erotetaan eri arvot yleisellä valuuttasymbolilla "¤"
  - Dokumentilla voi olla seuraavia attribuutteja

    1: Tekijät

    2: Nimeke

    3: Julkaisija

    4: Julkaisu

    5: JulkaisuVuosi

    6: Viittaustunniste

    7: Viitatut

    8: Viittaavat

    9: Hakusanat

Liite 2: Nimi -tyyppisen pikaformaattidatan määritelmä

- Merkistönä UTF-8, rivinerotin LF-merkki
- #-merkillä alkava rivi on kommenttirivi
- Kaikki tyhjät ja kommenttirivit jätetään huomiotta
- Ensimmäisellä ei-kommenttirivillä on tietueiden tyyppi:

    2=Toimija, tyyppi tuntematon

    2.1=Henkilö

    2.2=Organisaatio

    3=Rooli, tyyppi tuntematon

    3.1=Julkaisija

    4=Kanava, tyyppi tuntematon

    4.1=Lehti

    4.2=Konferenssi

    4.3=Julkaisusarja

    4.4=Raporttisarja

    4.5=Tietokanta
- Toisella ei-kommenttirivillä on lähteen nimi
- Joka tietue koostuu yhdestä rivistä, jolla on ensin resurssin nimi (Resurssin nimi on lähdekohtainen yksikäsitteinen tunniste)
- Jos resurssilla on useita vaihtoehtoisia nimiä, erotetaan eri arvot yleisellä valuuttamerkillä ¤
- Nimen jälkeen optionaalisesti kaksoispiste ja välilyönti(:), jonka jälkeen resurssia (l. roolia) "hallinnoiva" toimija

    Esimerkki tapauksesta, jossa julkaisijalla 2 vaihtoehtoista nimeä:
    Springer¤Springer-Verlag

    Esimerkki tapauksesta, jossa on ilmoitettu julkaisijaa hallinoiva organisaatio:
    Microsoft Press: Microsoft
- Lopussa optionaalinen pystyviiva | jonka jälkeen enintään kolme pilkuilla erotettua hakusanaa