# Vaatimusdokumentti

Boa Open Access

Helsinki 31.3.2006

Ohjelmistotuotantoprojekti

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

**Kurssi**

581260 Ohjelmistotuotantoprojekti (6 ov)

**Projektiryhmä**

Ilmari Heikkinen
Timo Hintsa
Erno Härkönen
Arto Vuori
Mikko Kautto

**Asiakas**

Olli Niinivaara

**Johtoryhmä**

Juha Taina
Riikka Kaven

**Kotisivu**

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

**Versiohistoria**

| Versio | Päiväys | Tehdyt muutokset |
|---|---|---|
| 1.0 | 27.1.2006 | Alustava versio |
| 1.4 | 5.2.2006 | Vaatimukset rajattu ensimmï¿œstï¿œiteraatiota koskeviin |
| 1.5 | 9.2.2006 | Vaatimukset korjattu palautteen pohjalta |
| 1.6 | 13.2.2006 | Vaatimuksia muutettu palautteen pohjalta |
| 1.7 | 27.3.2006 | Toisen iteraation vaatimukset |

# Sisältö

**1 Johdanto** **1**

**2 Sanasto** **1**

**3 Prioriteetit** **2**

**4 Käyttäjävaatimukset** **2**

4.1 Qriterium Statement (QS) -tiedostomuodon määrittely . . . . . 2

4.1.1 QS-muodon DTD . . . . . 2

4.1.2 OAI Dublin Coren määrittämä osajoukko . . . . . 2

4.2 XML Dublin Core -muunnos . . . . . 2

4.3 Käyttäjän tietojen syöttäminen . . . . . 3

4.4 Raakadatalähteen tietojen määrittely . . . . . 3

4.5 DBLP-muunnos . . . . . 3

4.6 Qriterium Statement -muunnos . . . . . 3

4.7 Helposti laajennettava arkkitehtuuri . . . . . 3

4.7.1 QS-muodon alustariippumattomuus . . . . . 3

4.7.2 Kirjaston uudelleenkäytettävyys . . . . . 3

4.7.3 QS-muodon määrittely . . . . . 4

4.7.4 QS-muodon katselu selaimella . . . . . 4

4.7.5 Muuntotyökalujen uudelleenkäytettävyys . . . . . 4

4.8 Vakaus . . . . . 4

4.8.1 Ei-validin raakadatan käsittely . . . . . 4

4.8.2 Tuntemattomat määritteet . . . . . 4

4.8.3 Virheellisten tietueiden kerääminen omaan tiedostoonsa . . . . . 5

4.9 Käyttöliittymä . . . . . 5

4.9.1 Virheilmoitukset . . . . . 5

4.9.2 Ajonaikainen palaute . . . . . 5

4.9.3 Ulostulotiedostojen koon määrittely . . . . . 5

4.9.4 Mahdollisuus pysäyttää eräajo siististi . . . . . 5

4.9.5 Mahdollisuus jatkaa pysäytettyä eräajoa . . . . . 5

4.9.6 Lisätyn metadatan poisto . . . . . 6

**5 Käyttötapaukset** **6**

5.1 Raakadatan muunnos levyllä olevasta XML Dublin Core -muotoisesta raakadatatiedostosta QS-muotoon . . . . . 6

5.2 Raakadatan muunnos levyllä olevasta DBLP-muotoisesta raakadatatiedostosta QS-muotoon . . . . . 6

5.3 QS-muotoisen tiedoston muuntaminen QS-muotoiseksi tiedostoksi . . . . 6

5.4 Käyttäjätietojen määrittäminen . . . . . 6

5.5 Raakadatalähteen tietojen määrittäminen . . . . . 6

**6 Järjestelmävaatimukset** **7**

6.1 Yhteensopivuus . . . . . 7

6.2 Laitevaatimukset . . . . . 7

6.3 Käytettävät komponentit . . . . . 7

**7 Ympäristövaatimukset** **7**

7.1 Käsiteltävien tiedostojen koko . . . . . 7

7.2 Suorituskyky . . . . . 7

**8 Järjestelmäarkkitehtuuri** **7**

8.1 Muunnosrajapinta . . . . . 7

8.2 Vientirajapinta . . . . . 8

8.3 Tiedonlukurajapinta . . . . . 8

**9 Qriterium Statement 0.2.4 (2006-02-13)** **8**

9.1 QID . . . . . 8

9.2 HEADER . . . . . 8

9.3 RESOURCE . . . . . 9

9.4 FROMCONNECTIONS . . . . . 9

9.5 TOCONNECTIONS . . . . . 9

9.6 Tyyppikohtaiset attribuutit . . . . . 10

9.6.1 TYPE: ACTOR . . . . . 10

9.6.2 TYPE: DOCUMENT . . . . . 10

9.6.3 TYPE: DATASOURCE . . . . . 11

9.6.4 TYPE: CONTENT . . . . . 11

**10 Esimerkkimuunnos CiteSeerin raakadatasta QS:ksi** **12**

10.1 Alkuperäinen CiteSeer-raakadata . . . . . 12
10.2 CiteSeer-raakadata muunnettuna QS-muotoiseksi XML:ksi . . . . . 14

# 1 Johdanto

Tämä on ohjelmistotuotantoprojektiryhmän Boa Open Access (BOA) vaatimusdokumentti.

Open Access tarkoittaa tieteellisen tiedon julkaiseminen internetissä siten, että se on vapaasti kenen tahansa luettavissa, tulostettavissa ja levitettävissä edelleen ilman maksuja tai käytön esteitä.

Tämän pilottiprojektin tarkoituksena on aloittaa Open Access-viittaustietokannan toteuttaminen muunnostyökalusta, joka muuttaa erilaisten metadatatiedostojen tiedostomuodon myöhemmin toteutettavan muun järjestelmän ymmärtämään muotoon.

BOA-projekti noudattaa Open Access -ideologiaa: siinä käsitellään Open Access -tietoa, tuotettu metadata on avoimesti saatavilla, ja ohjelmistoarkkitehtuuri on avoimen lisenssin alainen.

Projektissa tuotettujen speksien, ohjelmien käyttöliittymän, ohjelmakoodin ja koodin kommenttien kielenä käytetään englantia jatkokehitystä silmälläpitäen. Projektissa tuotetut sisäiset dokumentit vaatimusdokumentti mukaanlukien ovat suomenkielisiä.

# 2 Sanasto

**Metadata** Tieto, joka kuvaa muuta tietoa.

**Raakadata (raw data)** Palvelimella tai paikallisessa tietovarastossa sijaitseva metadata, jota muunnostyökalu käsittelee.

**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. Ohjelman sisäinen muoto.

**Qriterium Statement (QS) -formaatti; QS-muoto** Metadatatiedostomuoto, joka sisältää myöhemmin määritellyt Qriterium Statementin määrittelemät tiedot.

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

**DBLP** "Digital Bibliography & Library Project."Tietojenkäsittelytieteen julkaisujen bibliografia Internetissä. Sisältää yli 700.000:n julkaisun tiedot (alkuvuodesta 2006.) URL: http://dblp.uni-trier.de/

**Transformointi (muunnos, transformation)** Prosessi jossa eri tyyppisiä metadataformaatteja muunnetaan yhteiseen atomilauseformaattiin.

**Muunnostyökalu** Ohjelma, joka suorittaa ylläkuvattua muunnosprosessia.

**QS-kirjasto** Ohjelmointikirjasto, joka on suunniteltu QS-muotoisten tiedostojen käsittelyyn.

**Otsaketiedot** QS-muotoisessa metadatassa oleva tietue, joka kertoo kuka sen on tehnyt, milloin, ja miten.

**MAITO-projekti** Syksyllä 2005 toiminut ohjelmistotuotantoprojektiryhmä, joka teki ohjelmistonsa jotakuinkin samasta aiheesta: http://www.cs.helsinki.fi/group/metadata/webfinal/

# 3 Prioriteetit

Vaatimuksissa käytettyjen prioriteettitasojen kuvaukset:

**1** Kriittinen vaatimus. Korkein prioriteetti, projektin onnistumiselle välttämätön.

**2** Tärkeä vaatimus. Toteutetaan, jos 1. luokan prioriteettien jälkeen jää aikaa. Projektissa tavoiteltavan laadun saavuttamiseksi tärkeä.

**3** Parannusvaatimus. Toteutetaan, jos 1. ja 2. luokan prioriteetit jo toteutettu. Parantaa tuotettavan ohjelmiston tai dokumentin toiminnallisuutta.

**Ei toteuteta** Vaatimus, jonka toteuttaminen ei mahdu projektin aikatauluun.

# 4 Käyttäjävaatimukset

## 4.1 Qriterium Statement (QS) -tiedostomuodon määrittely

### 4.1.1 QS-muodon DTD

XML DTD:n tuottaminen QS-muodon syntaksista. DTD:tä käyttäen syntaktisesti korrektin QS-muotoisen tiedoston lukeminen ja kirjoittaminen pitäisi olla mahdollista.
Prioriteetti: 1 (toteutettu, iteraatio 1)

### 4.1.2 OAI Dublin Coren määrittämä osajoukko

Suurin prioriteetti QS-muodon toteuttamisessa on OAI Dublin Core -muotoisen metadatan sisältämän informaation ilmaisemisessa.
Prioriteetti: 1 (toteutettu, iteraatio 1)

## 4.2 XML Dublin Core -muunnos

Käyttäjä voi muuntaa XML Dublin Core -muotoisen raakadatatiedoston QS-muotoisiksi tiedostoiksi. Muunnoksen semantiikat seurailevat MAITO-projektin vastaavia.
Prioriteetti: 1 (toteutettu, iteraatio 1)

## 4.3 Käyttäjän tietojen syöttäminen

Käyttäjä voi syöttää tiedot itsestään jollain tavalla, esimerkiksi tekemällä QS-muotoisen tiedoston, josta tiedot luetaan.
Prioriteetti: 1 (iteraatio 2)

## 4.4 Raakadatalähteen tietojen määrittely

Käyttäjä voi syöttää tiedot syötettävän raakadatan lähteestä jollain tavalla, esimerkiksi tekemällä QS-muotoisen tiedoston, josta tiedot luetaan.
Prioriteetti: 1 (iteraatio 2)

## 4.5 DBLP-muunnos

Käyttäjä voi muuntaa DBLP-muotoisen raakadatatiedoston QS-muotoisiksi tiedostoiksi.
Prioriteetti: 1 (iteraatio 2)

## 4.6 Qriterium Statement -muunnos

Käyttäjä voi muuntaa QS-muotoisen tiedoston uusiksi QS-muotoisiksi tiedostoiksi. Näin tehtäessä id:t päivitetään ja käyttäjän tiedot lisätään.
Prioriteetti: 1 (iteraatio 2)

## 4.7 Helposti laajennettava arkkitehtuuri

Jatkokehittäjät voivat kirjoittaa uusia muuntimia muunninkirjastoa hyväksikäyttäen, tai kirjoittamalla kokonaan uusia muuntimia tarvitsematta käyttää BOA-projektissa kehitettyjä työkaluja.
Prioriteetti: 1 (jatkuva, toteutettu 1. iteraation osalta)

### 4.7.1 QS-muodon alustariippumattomuus

QS-muodon ei pidä riippua BOA-projektissa toteutetuista työkaluista ja kirjastoista.
Prioriteetti: 1 (toteutettu, iteraatio 1)

### 4.7.2 Kirjaston uudelleenkäytettävyys

Ohjelman QS-kirjaston tulee olla seuraavan projektiryhmän helposti uudelleenkäytettävissä, niin ettei jatkokehittäjien tarvitse toteuttaa uudestaan jo toteutettuja osia.
Prioriteetti: 2 (jatkuva, toteutettu 1. iteraation osalta)

### 4.7.3 QS-muodon määrittely

QS-muodon määritelmän tulee olla sellainen, että sitä kirjoittavien ja lukevien työkalujen kirjoittaminen on helppoa.
Prioriteetti: 2 (toteutettu, iteraatio 1)

### 4.7.4 QS-muodon katselu selaimella

QS-muodon tulee olla sellainen, että sitä voi katsella WWW-selaimella. Vaihtoehtoisesti tulisi toteuttaa komponentti, joka muuntaa QS-muotoista dataa selaimella katseltavaan muotoon (esim. XSLT:tä käyttäen.)
Prioriteetti: 3 (iteraatio 3)

### 4.7.5 Muuntotyökalujen uudelleenkäytettävyys

Muuntotyökalujen tulee olla mahdollisimman riippumattomia toisistaan, niin että johonkin työkaluun tehdyt muutokset eivät aiheuta tarvetta tehdä muutoksia muihin työkaluihin.
Prioriteetti: 2 (jatkuva, toteutettu 1. iteraation osalta)

## 4.8 Vakaus

Ohjelma ei saa kaatua virheellisellä syötteellä.
Prioriteetti: 1 (jatkuva, toteutettu 1. iteraation osalta)

### 4.8.1 Ei-validin raakadatan käsittely

Jos parsittava (XML-)raakadata ei ole validi rakenteeltaan, tähän asti tuotettu tulostiedosto kirjoitetaan levylle ja loput raakadatasta ohitetaan. Jos virheestä toipuminen on mahdollista (esim. tiedetään tietueen aloittavan tagin nimi), niin yritetään mahdollisuuksien mukaan ohittaa vain virheellinen kohta ja jatkaa raakadatan käsittelyä.
Prioriteetti: 1 (jatkuva, toteutettu 1. iteraation osalta)

### 4.8.2 Tuntemattomat määritteet

Tuntemattomat määritteet (esim. XML-tagit) ohitetaan siirtymällä jäsentämään määritettä seuraavaa kohtaa dokumentissa.
Prioriteetti: 1 (jatkuva, toteutettu 1. iteraation osalta)

### 4.8.3 Virheellisten tietueiden kerääminen omaan tiedostoonsa

Virheelliset tietueet kerätään omaan tiedostoonsa. Tarkoituksena helpottaa virheiden havaitsemista ja korjausta tarvitsematta käydä koko alkuperäistä syötedataa läpi.
Prioriteetti: 2 (iteraatio 2)

## 4.9 Käyttöliittymä

### 4.9.1 Virheilmoitukset

Ohjelma kertoo käyttäjälle mahdollisista virhetilanteista. Virheen tapahtuessa kerrotaan käyttäjälle virheen tapahtumisesta, syystä, ja, parsintavirheen kyseessä ollessa, jäsennettävän dokumentin nimi sekä rivi numeroineen, jolla virhe tapahtui.
Prioriteetti: 1 (iteraatio 2)

### 4.9.2 Ajonaikainen palaute

Ohjelma antaa käyttäjälle mielekästä palautetta parhaillaan meneillään olevasta operaatiosta. Vähintään mitä dokumenttia parhaillaan käsitellään, ja kuinka suuri määrä siitä on jo käsitelty.
Prioriteetti: 2 (iteraatio 2)

### 4.9.3 Ulostulotiedostojen koon määrittely

Tuotettavien QS-tiedostojen maksimikoko tulee olla määritettävissä tavuissa.
Prioriteetti: 3 (toteutettu 1. iteraation osalta)

### 4.9.4 Mahdollisuus pysäyttää eräajo siististi

Käyttäjä voi pysäyttää ohjelman suorituksen niin, ettei jo luotu QS-muotoinen tiedosto häviä. Lisäksi ohjelman tulee toimia oikein seuraavalla ajokerralla niin, että esimerkiksi seuraavan ajokerran id:t eivät konfliktoi keskeytetyn ajon tulostiedostojen kanssa.
Prioriteetti: 2 (iteraatio 2)

### 4.9.5 Mahdollisuus jatkaa pysäytettyä eräajoa

Käyttäjä voi jatkaa pysäytettyä eräajoa.
Prioriteetti: 3 (iteraatio 2)

#### 4.9.6 Lisätyn metadatan poisto

Käyttäjä voi poistaa järjestelmään lisättyä metadataa, esim. poistamalla luodun QS-muotoisen tiedoston.
Prioriteetti: 3 (iteraatio 2)

# 5 Käyttötapaukset

## 5.1 Raakadatan muunnos levyllä olevasta XML Dublin Core -muotoisesta raakadatatiedostosta QS-muotoon

Käyttäjä muuntaa levyllä olevan XML Dublin Core -muotoisen raakadatatiedoston QS-muotoikseksi tiedostoksi.
Prioriteetti: 1 (toteutettu, iteraatio 1)

## 5.2 Raakadatan muunnos levyllä olevasta DBLP-muotoisesta raakadatatiedostosta QS-muotoon

Käyttäjä muuntaa levyllä olevan DBLP-muotoisen raakadatatiedoston QS-muotoiseksi tiedostoksi.
Prioriteetti: 1 (iteraatio 2)

## 5.3 QS-muotoisen tiedoston muuntaminen QS-muotoiseksi tiedostoksi

Käyttäjä muuntaa levyllä olevan QS-muotoisen tiedoston uusiksi QS-muotoisiksi tiedostoiksi, joiden id:t on päivitetty ja käyttäjän tiedot lisätty.
Prioriteetti: 1 (iteraatio 2)

## 5.4 Käyttäjätietojen määrittäminen

Käyttäjä kertoo ohjelmalle tietonsa tulevien eräajojen otsaketietoja varten.
Prioriteetti: 1 (iteraatio 2)

## 5.5 Raakadatalähteen tietojen määrittäminen

Käyttäjä kertoo ohjelmalle raakadatan lähteen tiedot raakadatasta tehtävän QS-muotoisen tiedoston otsaketietoja varten.
Prioriteetti: 1 (iteraatio 2)

# 6 Järjestelmävaatimukset

## 6.1 Yhteensopivuus

Muunnosohjelma toimii millä tahansa tietokoneella, jolla on Java-tulkki (versio >= 1.5.)
Prioriteetti: 1 (toteutettu iteraation 1 puitteissa)

## 6.2 Laitevaatimukset

Ohjelma toimii 256 megatavulla muistia ja vaatii (5 + tf) kertaa syötettävän datan verran vapaata levytilaa, jossa tf on tulosmuodon koko jaettuna syötettävän datan koolla.
Prioriteetti: 3 (iteraatio 3)

## 6.3 Käytettävät komponentit

Kaikkien muunnosohjelmassa käytettyjen komponenttien on oltava vapaan ohjelmistolisenssin alaisia tai ilmaiseksi saatavilla.
Prioriteetti: 1 (toteutettu iteraation 1 puitteissa)

# 7 Ympäristövaatimukset

## 7.1 Käsiteltävien tiedostojen koko

Ohjelman pitää kyetä käsittelemään yli kahden gigatavun kokoisia syötetiedostoja.
Prioriteetti: 1 (toteutettu iteraation 1 puitteissa)

## 7.2 Suorituskyky

Ohjelma pystyy käsittelemään vähintään 100 megatavua syötettä tunnissa.
Prioriteetti: 1 (toteutettu iteraation 1 puitteissa)

# 8 Järjestelmäarkkitehtuuri

Muunnosohjelmisto koostuu kolmesta osasta.

## 8.1 Muunnosrajapinta

Muunnosrajapinnan toteuttavat komponentit lukevat raakadataa ja muuntavat sitä ohjelman sisäiseen atomilausemuotoon.

## 8.2 Vientirajapinta

Vientirajapinnan toteuttavat komponentit muuntavat ohjelman sisäistä atomilausemuotoista dataa QS-muotoisiksi tiedostoiksi.

## 8.3 Tiedonlukurajapinta

Yhdistää erilaiset tiedonlukuprotokollat yhtenäisen rajapinnan alle, tarkoituksena tarjota varsinaisille raakadatamuuntimille siirtomuodosta itsenäinen tapa lukea raakadataa.

# 9 Qriterium Statement 0.2.4 (2006-02-13)

Qriterium Statement määrittää informaation, jonka QS-tiedostomuodon pitäisi kyetä ilmaisemaan, varsinainen tiedostomuodon tarkempi määritelmä on osa projektin toteutusvaihetta.

Tavoitteena on kirjoittaa auki raakadatan sisältämä implisiittinen suhdeverkko erilaisten resurssien välillä.

Tuotettava data on UTF-8-enkoodattua.

Spesifikaatiossa voidaan antaa ohjeistusta kenttien koosta.

Jokaisessa lauseessa on Header ja välissä jokin seuraavista:

RESOURCE

FROMCONNECTIONS

TOCONNECTIONS

## 9.1 QID

Resurssin tunnistemerkkijono. Mahdollisimman globaalisti uniikki.

## 9.2 HEADER

SIGN
- Statementin digitaalinen allekirjoitus
- Samalla statementin ja siinä mahd. esiteltävän resurssin globaali yksikäsitteinen tunniste

SS - Tämän statement sourcen QID

BY - Statementin tekijän QID

SSNR -Statementin SS-kohtainen juokseva numero

BYNR - Statementin BY-kohtainen juokseva numero

ORIGIN - lauseen tietolähde (datalähteen, dokumentin tai BY:n QID)

AT - Statementin luonnin pvm ja kellonaika (ISO-TIME)

## 9.3 RESOURCE

QID - Määrittely edellä

TYPE - Resurssin tyyppi: ACTOR, DOCUMENT, DATASOURCE tai CONTENT

SUBTYPE - Resurssin alityyppi, riippuu tyypistä

## 9.4 FROMCONNECTIONS

FROM - Yhteyden alkupään (=uudemman) resurssin QID

## 9.5 TOCONNECTIONS

TO - Yhteyden loppupään (=vanhemman) resurssin QID

FROMCONNECTION ja TOCONNECTION sisältävät yhden tai useamman yhteyden seuraavasta listasta. Jokaiseen yhteyteen liittyy QID. Voisi myös liittyä NOT (NOT CONTACTED_AT jne.).

CONTACTED_AT - Yhteystieto, kuten organisaatio tai sähköposti tai molemmat.

MANAGED_BY - Hallintosuhde, kuten henkilö töissä organisaatiossa.

STORED_BY - Tallentaja

PUBLISHED_BY - Julkaisija

IDENTIFIED_AS - Tämän resurssin vaihtoehtoinen QID

PRESENTED_BY - Esittäjä, esim. kääntäjä

COMPOSED_OF - Koostumus

CREATED_BY - Tekijä

CREATED_AT - Luontipaikka

CREATED_FOR - Tilaaja

SUPPORTED_BY - Tukija

AWARE_OF - Viitatut aiheet

## 9.6 Tyyppikohtaiset attribuutit

### 9.6.1 TYPE: ACTOR

ACTOR määrittelee toimijan. Toimija on nimetty entiteetti.

SUBTYPE - Toimijan alityyppi: PERSON, ORGANIZATION, ANONYMOUS (anonyymi käyttäjä, jolla ei omaa "avainta"), UNKNOWN

NAME - Koko nimi, henkilönimet muodossa "sukunimi, etunimi".
Esim. International Business Machines Corporation.

PRIMARYNAME - Sukunimi tai esim. IBM

SECONDARYNAMES - Etunimi tai vrt. Microsoft<->MS

NAMEEXTENSIONS - Ylimääräinen nimen määrite, esim. Mr., Jr., Corp.

NICKNAME - Lempinimi, esim. Big Blue

RAWNAME - Alkuperäinen raakadatassa ollut nimi

### 9.6.2 TYPE: DOCUMENT

DOCUMENT määrittelee dokumentin metadatan, sisällön metadata on CONTENTissa.

SUBTYPE - Dokumentin alatyyppi, UNKNOWN

TITLE - Dokumentin nimi

RAWNAME - Alkuperäinen raakadatassa ollut nimi

LANGUAGE - Dokumentissa käytetty kieli (ISO 639)

KEYWORDS - Dokumenttiin liittyvä avainsanat

CREATED - Dokumentin luontipäivämäärä (ISO-TIME)

PUBLISHED - Dokumentin julkaisupäivämäärä (ISO-TIME)

BIBCIT - Bibliografinen viittaus, esim. Marx, K., Das Kapital, vol. 1, Hamburg, Germany: Meissner, 1867. Sisällytetään sellaisenaan, jos löytyy raakadatasta.

OPENURL - OpenURL-viittaus (http://openurl.info/registry). Sisällytetään sellaisenaan, jos löytyy raakadatasta.

### 9.6.3 TYPE: DATASOURCE

DATASOURCE määrittää raakadatatietolähteen.

SUBTYPE - Datalähteen alatyyppi: DOCUMENT, METADATA, STATEMENT, NET, CONTENT, RANKING, UNKNOWN

NAME - Datalähteen nimi

ENCODING - Datalähteen käyttämä tekstienkoodaus.
Esim. UTF-8

FORMAT - Datalähteen käyttämä MIME-tyyppi.
Esim. text/plain, application/xml, application/pdf

TERMINOLOGY - Datalähteen käyttämä ontologia.
Esim. Dublin Core, MODS, MARC

COMPRESSION - Datalähteen mahdollisesti käyttämän tiedoston pakkausmekanismin MIME-tyyppi.
Esim. application/x-compressed-tar, application/x-bzip-compressed-tar

BACKPROTOCOL - Datalähteen alempi sovellustason tiedonsiirtoprotokolla.
Esim. SMTP, HTTP, Freenet, Bittorrent

TOPPROTOCOL - Datalähteen ylempi sovellustason tiedonsiirtoprotokolla.
Esim. Qriterium, OAI-PMH, OpenURL, SRU

LAUNCHED - Päivämäärä, josta lähtien datalähde on kerännyt dataa.
Esim. UNKNOWN

### 9.6.4 TYPE: CONTENT

CONTENT määrittää aiheet, mitä dokumentti käsittelee

SUBTYPE - Sisällön alatyyppi: UNKNOWN

KEYWORDS - Kaikkien dokumenttien KEYWORDS

CREATED - Ensimmäisen tunnetun dokumentin PUBLISHED

NAME - Ensimmäisen tunnetun dokumentin nimi.

# 10 Esimerkkimuunnos CiteSeerin raakadatasta QS:ksi

## 10.1 Alkuperäinen CiteSeer-raakadata

```
<record>
<header>
<identifier>oai:CiteSeerPSU:10</identifier>
<datestamp>1998-07-02</datestamp>
<setSpec>CiteSeerPSUset</setSpec>
</header>
<metadata>
<oai_citeseer:oai_citeseer
xmlns:oai_citeseer="http://copper.ist.psu.edu/oai/
oai_citeseer/" xmlns:dc
="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://copper.ist.psu.edu/oai/
oai_citeseer/
http://copper.ist.psu.edu/oai/oai_citeseer.xsd ">
   <dc:title>Optimized Software Synthesis for Digital Signal
Processing
Algorithms -- An Evolutionary Approach</dc:title>
   <oai_citeseer:author name="Jurgen Teich">
      <address>Gloriastrasse 35 , CH-8092 Zurich;
Switzerland</address>
      <affiliation>Computer Engineering and Communication
Networks Lab (TIK);
Swiss Federal Institute of Technology (ETH)</affiliation>
   </oai_citeseer:author>
   <oai_citeseer:author name="Eckart Zitzler">
      <address>Gloriastrasse 35 , CH-8092 Zurich;
Switzerland</address>
      <affiliation>Computer Engineering and Communication
Networks Lab (TIK);
Swiss Federal Institute of Technology (ETH)</affiliation>
   </oai_citeseer:author>
   <oai_citeseer:author name="Shuvra S. Bhattacharyya">
      <address>College Park MD 20742</address>
      <affiliation>Department of Electrical Engineering ,
and; Institute for
Advanced Computer Studies (UMIACS); University of
Maryland</affiliation>
   </oai_citeseer:author>
   <dc:subject>Jurgen Teich,Eckart Zitzler,Shuvra S.
```

Bhattacharyya Optimized
Software Synthesis for Digital Signal Processing Algorithms
-- An Evolutionary
Approach</dc:subject>
<dc:description>This paper addresses the problem of
trading-off between the
minimization of program
and data memory requirements of single-processor
implementations of dataflow
programs. Based on the formal model of synchronous data flow
(SDF) graphs
[ LM87 ] , so called single appearance schedules are known
to be program-memory
optimal. Among these schedules, buffer memory schedules are
investigated and
explored based on a two-step approach: (1) An Evolutionary
Algorithm (EA) is
applied to efficiently explore the (in general) exponential
search space of
actor firing
orders. (2) For each order, the buffer costs are evaluated
by applying a dynamic
programming
post-optimization step (GDPPO). This iterative approach is
compared
to existing heuristics for buffer memory optimization.
Chapter 1
Introduction
Dataflow specifications are widespread in areas of digital
signal and image
processing.
In dataflow, a specification consists of a directed graph in
which the nodes
represent computations and the arcs ...</dc:description>
<dc:contributor>The Pennsylvania State University
CiteSeer
Archives</dc:contributor>
<dc:publisher>unknown</dc:publisher>
<dc:date>1998-07-02</dc:date>
<oai_citeseer:pubyear>1998</oai_citeseer:pubyear>
<dc:format>ps</dc:format>

<dc:identifier>http://citeseer.ist.psu.edu/10.html</dc:
identifier>

```
<dc:source>ftp://ftp.tik.ee.ethz.ch/pub/people/zitzler/
TZB1998a.ps.gz</dc:source
>
   <dc:language>en</dc:language>
   <oai_citeseer:relation type="References">

<oai_citeseer:uri>oai:CiteSeerPSU:62368</oai_citeseer:uri>
   </oai_citeseer:relation>
   <oai_citeseer:relation type="References">

<oai_citeseer:uri>oai:CiteSeerPSU:185918</oai_citeseer:uri>
   </oai_citeseer:relation>
   <oai_citeseer:relation type="References">

<oai_citeseer:uri>oai:CiteSeerPSU:24284</oai_citeseer:uri>
   </oai_citeseer:relation>
   <oai_citeseer:relation type="References">

<oai_citeseer:uri>oai:CiteSeerPSU:5184</oai_citeseer:uri>
   </oai_citeseer:relation>
   <oai_citeseer:relation type="References">

<oai_citeseer:uri>oai:CiteSeerPSU:121879</oai_citeseer:uri>
   </oai_citeseer:relation>
   <oai_citeseer:relation type="References">

<oai_citeseer:uri>oai:CiteSeerPSU:22987</oai_citeseer:uri>
   </oai_citeseer:relation>
   <dc:rights>unrestricted</dc:rights>
</oai_citeseer:oai_citeseer>
</metadata>
</record>
```

## 10.2 CiteSeer-raakadata muunnettuna QS-muotoiseksi XML:ksi

```
-- vanhentunut, korvaa uudella --

<qs:statement_list>


<qs:statement>

<qs:header>
```

```
  <qs:id>123</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>6</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>44</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:document>
  <qs:title>Optimized Software Synthesis for Digital Signal
Processing
Algorithms -- An Evolutionary Approach</qs:title>
  <qs:language>en</qs:language>
  <qs:published>1998-07-02</qs:published>

<qs:openurl>http://citeseer.ist.psu.edu/10.html</qs:openurl>
</qs:document>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>124</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>7</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>45</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:actor subtype="person">
  <qs:fullname>Jurgen Teich</qs:fullname>
</qs:actor>

</qs:statement>


<qs:statement>
```

```
<qs:header>
  <qs:id>125</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>8</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>46</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:actor subtype="person">
  <qs:fullname>Eckart Zitzler</qs:fullname>
</qs:actor>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>126</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>9</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>47</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:actor subtype="person">
  <qs:fullname>Shuvra S. Bhattacharyya</qs:fullname>
</qs:actor>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>127</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>10</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
```

```
  <qs:origin>13</qs:origin>
  <qs:ssnr>48</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:actor subtype="organization">
  <qs:fullname>Computer Engineering and Communication
Networks Lab(TIK); Swiss
Federal Institute of Technology (ETH)</qs:fullname>
</qs:actor>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>128</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>11</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>49</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:actor subtype="organization">
  <qs:fullname>Department of Electrical Engineering, and ;
Institute for
Advanced Computer Studies (UMIACS); University of
Maryland</qs:fullname>
</qs:actor>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>129</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>12</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
```

```
  <qs:origin>13</qs:origin>
  <qs:ssnr>50</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:content>
</qs:content>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>130</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>13</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>51</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:actor subtype="unknown">
</qs:actor>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>131</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>14</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>52</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="created_by">
```

```
  <qs:from_id>123</qs:from_id>
  <qs:to_id>124</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>132</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>15</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>53</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="created_by">
  <qs:from_id>123</qs:from_id>
  <qs:to_id>125</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>133</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>16</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>54</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="created_by">
  <qs:from_id>123</qs:from_id>
  <qs:to_id>126</qs:to_id>
</qs:connection>
```

```
</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>134</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>17</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>55</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="published_by">
  <qs:from_id>123</qs:from_id>
  <qs:to_id>130</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>135</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>18</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>56</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="managed_by">
  <qs:from_id>124</qs:from_id>
  <qs:to_id>127</qs:to_id>
</qs:connection>

</qs:statement>
```

```
<qs:statement>

<qs:header>
  <qs:id>136</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>19</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>57</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="managed_by">
  <qs:from_id>125</qs:from_id>
  <qs:to_id>127</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>137</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>20</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>58</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="managed_by">
  <qs:from_id>126</qs:from_id>
  <qs:to_id>128</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>
```

```
<qs:header>
  <qs:id>138</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>21</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>59</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="consists_of">
  <qs:from_id>123</qs:from_id>
  <qs:to_id>129</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>139</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>22</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>60</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="aware_of">
  <qs:from_id>129</qs:from_id>
  <qs:to_id>79</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>140</qs:id>
  <qs:by>65</qs:by>
```

```
  <qs:bynr>23</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>61</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="aware_of">
  <qs:from_id>129</qs:from_id>
  <qs:to_id>65</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>141</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>24</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>62</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="aware_of">
  <qs:from_id>129</qs:from_id>
  <qs:to_id>81</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>142</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>25</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
```

```
  <qs:ssnr>63</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="aware_of">
  <qs:from_id>129</qs:from_id>
  <qs:to_id>11</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>143</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>26</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>64</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
</qs:header>

<qs:connection type="aware_of">
  <qs:from_id>129</qs:from_id>
  <qs:to_id>9</qs:to_id>
</qs:connection>

</qs:statement>


<qs:statement>

<qs:header>
  <qs:id>143</qs:id>
  <qs:by>65</qs:by>
  <qs:bynr>26</qs:bynr>
  <qs:rawid>oai:CiteSeerPSU:10</qs:rawid>
  <qs:origin>13</qs:origin>
  <qs:ssnr>64</qs:ssnr>
  <qs:cat>1998-07-02</qs:cat>
  <qs:datestamp>2006-02-19</qs:datestamp>
```

```
</qs:header>

<qs:connection type="stored_by">
  <qs:from_id>123</qs:from_id>
  <qs:to_id>13</qs:to_id>
</qs:connection>

</qs:statement>


</qs:statement_list>
```