|  |
| --- |
| Dansk Data Arkiv |
| DDI indekseringsplatform og søgemaskine |
| Implementeret som eXist DB med Lucene som indeks |

|  |
| --- |
| Kemal Pajevic  08-01-2012 |

Indhold

[Introduktion 2](#_Toc313819406)

# Introduktion

## Formål

Formålet med projektet er at opsætte en XML database indeholdende studiedata i DDI format. Der skal efterfølgende opbygges indekser over udvalgte elementer i formatet for at effektivisere søgningen i dem, og endelig skal der implementeres søgefunktioner for at udtrække ønsket data fra databasen på baggrund af søgetekst samt nogle beskrivende parametre.

## eXist DB og Lucene

Til indekseringen af data har man valgt Lucene[[1]](#footnote-1) og eXist DB[[2]](#footnote-2) blev valgt som databasesystem. De er begge implementeret i Java og eXist bruger Lucene som sin default indekser. Interaktionen til eXist foregår igennem query sproget XQuery.

Den seneste udgivne version af eXist er 1.4, men denne indeholder en ældre version af Lucene som ikke inkluderer en såkaldt word analyzer til dansk. Derfor foregår udviklingen af dette projekt i en speciel branch af eXist projektet som findes på

<https://exist.svn.sourceforge.net/svnroot/exist/branches/adam/eXist-lucene-3.0>

En version 1.5 af eXist er i skrivende stund på trapperne, og denne bliver opdateret med den nyeste version af Lucene.

# Opbygning af projektet

## Filer og mapper

Projektmappen er delt op i følgende undermapper:

|  |  |  |  |
| --- | --- | --- | --- |
| Mappenavn | Indeholder | Del af kodebasen | Lægges  i eXist |
| denormalization | Filer til denormalisering af databasen | Ja | Ja |
| doc | Dokumentation af projektet som Word dokument samt html-baseret beskrivelse af moduler og funktioner (á la Java docs) | Nej | Nej |
| index | Konfigurationsfiler for at konfigurere indekserne | Ja | Ja |
| JavaRestClientDDA | Java testklient til test af projektets REST interface | Nej | Nej |
| lib | Moduler med søgefunktioner samt hjælpemoduler | Ja | Ja |
| rest | REST interfacet | Ja | Ja |
| schema | XSD skemaer til beskrivelse XML typer som anvendes i projektet. Bruges ikke af eXist | Ja | Nej |
| xml-sample-instances | Eksampel XML for typer anvendt i projektet | Nej | Nej |

Tabel - Oversigt over mapper i projektet

## Kodebasen

|  |  |  |
| --- | --- | --- |
| Filnavn | Beskrivelse | Beliggenhed i eXist |
| **denormalization** | | |
| denormalize.xquery | Opbygger denormaliseringsdata som bruges til optimering af søgningen | /db/dda-denormalization |
| **index** | | |
| dda/ collection.xconf | Indeks-konfigurationsfil til hovedcollection som indeholder DDI data | /db/system/config/db/dda |
| dda-denormalization/ collection.xconf | Indeks-konfigurationsfil til den denormaliserede collection | /db/system/config/db/ dda-denormalization |
| dda-urn/ collection.xconf | Indeks-konfigurationsfil til URN resolution collection | /db/system/config/db/ dda-urn |
| **lib** | | |
| result-functions.xquery | Funktioner til behandling og indpakning af søgeresultater | /db/dda/lib |
| search.xquery | Søgefunktioner | /db/dda/lib |
| urn.xquery | Funktioner til URN resolution | /db/dda/lib |
| **rest** | | |
| advanced-search.xquery | REST interface for avanceret søgning | /db/dda/rest |
| simple-search.xquery | REST interface for simpel søgning | /db/dda/rest |
| urn-resolution.xquery | REST interface for URN resolution | /db/dda/rest |
| **schema** | | |
| denormalization/ denormalized-ddi.xsd | Formatet af den denormaliserede collection | - |
| result/ ddieditor-lightxmlobject.xsd | Formatet som resultater returneres i | - |
| result/ result-metadata.xsd | Beskriver resultatet (antal hits, antal sider, aktuel side, osv.) | - |
| search/ advanced-search-parameters.xsd | Formatet af parameteret til avanceret søgning | - |
| search/ scope.xsd | Angiver ønskede DDI elementer (enten dem som ønskes fremsøgt eller dem som ønskes vist som referencer) | - |
| search/ search-metadata.xsd | Beskriver søgningen (ønsket antal resultater pr. side, anmodet side, osv.) | - |
| search/ simple-search-parameters.xsd | Formatet af parameteret til simpel søgning | - |

Tabel - Oversigt over filer som indgår i kodebasen

Mere detaljeret beskrivelse af moduler og funktioner kan findes i **doc** mappen.

# Opsætning

Der skal oprettes tre collections i roden (dvs. i den overordnede **/db** mappe) af eXist DB:

* **dda** - Her uploades DDI XML’erne. Deres placering i collectioen er underordnet, men det anbefales at de samles i en bestemt undercollection, f.eks. **data** eller **ddi** (navnet på dette undercollection er ligegyldig ift. søgningen).
* **dda-denormalization** - Her bliver det denormaliserede data gemt når denårmaliseringen eksekveres.
* **dda-urn** - Dette er en særskilt collection som også indeholder DDI data til URN resolution som kan indeholde flere versioner af DDI elementerne.

I indeks-konfigurationsmappen (**/db/system/config/db**) oprettes der tre tilsvarende collections hvor de tilhørende konfigurationsfiler (**collection.xconf**) lægges.

Filerne fra projektets kodebase uploades også til disse collections i de lokationer specificeret i Tabel 2. ”Undermapperne” i stierne oprettes som undercollections.

Husk at hvis indeks-konfigurationsfilerne lægges (eller ændres) efter at det data de bruges til at indeksere er blevet lagt ind så skal den relevante collection reindekseres. Data som er lagt efter at indekset er på plads indekseres automatisk.

Det er vigtigt at køre denormaliseringen efter data er lagt ind, for ellers findes ingen referencer ved søgningen.

# Denormalisering

For at nedsætte søgetiden er der implementeret en denormaliseret database som for samtlige relevante DDI elementer (Concept, Universe, QuestionItem, MultipleQuestionItem, Variable, Category) indeholder samtlige relevante referencer, dvs. elementer som enten refereres af eller referer til dette element.

For at opbygge denormaliseringsdata eksekveres scriptet **/db/dda-denormalization/denormalize.xquery**.

Denne proces sletter alt eksisterende data og genopbygger fra bunden.

Det er vigtigt at udføre processen med faste intervaller (f.eks. som skeduleret job som køres en gang om ugen) for at holde data opdateret. Den forventede (beregnet på baggrund af udførsel på mindre datasæt) køretid for processen med ca. 2500 studier i databasen er på ca. 2 timer.

For at optimere hastigheden af denne proces

1. [http://lucene.apache.org](http://lucene.apache.org/) [↑](#footnote-ref-1)
2. <http://exist-db.org/> [↑](#footnote-ref-2)