

# Fedora Software Afrunding

## Slutrapportering til DEFF Sekretariatet

Dato: 22. december 2011

### 1 Indhold

<b>2</b>	<b>Resumé af projektet</b>	<b>2</b>
<b>3</b>	<b>Generelle oplysninger</b>	<b>2</b>
<b>4</b>	<b>Økonomisk og juridisk ansvarlig</b>	<b>2</b>
<b>5</b>	<b>Samarbejdspartnere</b>	<b>3</b>
<b>6</b>	<b>Projektbeskrivelse</b>	<b>3</b>
6.1	<i>Baggrund</i>	3
6.2	<i>Mål</i>	4
6.3	<i>Målgruppe</i>	4
6.4	<i>Strategi/metode</i>	4
6.5	<i>Succeskriterier/resultater</i>	4
<b>7</b>	<b>Præsentation af projektet</b>	<b>5</b>
7.1	<i>Beskrivelse af projektforløbet</i>	5
7.2	<i>Beskrivelse og evaluering af projektets resultater</i>	5
7.2.1	Opdatering til de nyeste komponent-biblioteker	5
7.2.2	Videreudvikling af funktionaliteten på basis af opsamlede og prioriterede brugerønsker	5
7.2.3	Udvikling af reference-konfigurationer til støtte for implementeringen af nye Fedora-applikationer, inklusive en tutorial	6
7.2.4	Forbedring af performance, baseret på omfattende stress-tests	6
7.2.5	Formidling og support af GSearch via Fedora mail-listerne	6
<b>8</b>	<b>Konklusion</b>	<b>7</b>
<b>9</b>	<b>Links</b>	<b>7</b>
<b>10</b>	<b>Bilag</b>	<b>8</b>

## 2 Resumé af projektet

Projektet drejer sig om en fortsættelse og afrunding af de tidligere DEFF støttede projekter med Fedora Generic Search Service (kort GSearch), som blev inkluderet i Fedora Service Framework fra og med Fedoras version 2.2 i januar 2007.

Projektet har været meget vellykket. Det har produceret to nye Releases af GSearch som planlagt og præsenteret for Fedora samfundet. Overordnede og detaljerede forklaringer og vejledninger er indeholdt i OR2011-præsentationen og i downloadet. Fortløbende kommunikation med særligt interesserede brugere har medført flere nye features. Der er lagt op til yderligere kommunikation med brugerne på maillisten og direkte, og der vil efterfølgende blive fulgt op på eventuelle problemer og ønsker til nye features. Et paper om projektet vil blive indsendt til Open Repositories 2012 konferencen i Edinburgh i juli 2012.

Arbejdsindsatsen var 3 personmåneder.

Perioden var 1. juni – 31. december 2011.

2

## 3 Generelle oplysninger

### Projekt deltager:

#### Danmarks Tekniske Informationscenter, DTU Bibliotek

Projektleder/kontaktperson: Gert Schmeltz Pedersen, specialkonsulent, IT udvikling

#### Adresse:

Anker Engelunds Vej 101

DK-2800 Kgs. Lyngby

CVR-nr.: 30060946

Telefon: 4525 7244

Fax: 4588 3040

E-postadresse: [gsp@dtic.dtu.dk](mailto:gsp@dtic.dtu.dk)

Biblioteksnummer: 820060

## 4 Økonomisk og juridisk ansvarlig

#### Danmarks Tekniske Informationscenter

Direktør Mogens Sandfær

Anker Engelundsvej 1

Bygning 101D, rum 103B

2800 Kgs. Lyngby

Telefon: 4525 7311

E-postadresse: [ms@dtic.dtu.dk](mailto:ms@dtic.dtu.dk)

## 5 Samarbejdspartnere

Projektet er udført af Gert Schmeltz Pedersen, DTIC, i dialog med Fedora Committers, det centrale globale software team med 12 medlemmer

(<https://wiki.duraspace.org/display/FCREPO/Fedora+Contributors>), heraf 2 fra Danmark - i høj grad som følge af DEFF projekterne:

- Asger Askov Blekinge, Statsbiblioteket
- Gert Schmeltz Pedersen, DTIC.

## 6 Projektbeskrivelse

### 6.1 Baggrund

DEFF har i tidligere projekter [2][3] støttet dansk afprøvning af, implementering af og bidrag til Open Source softwaren Fedora Commons Repository System [4] som generisk repository software løsning for DEFF. I denne forbindelse har især Statsbiblioteket og DTIC været aktive i videreudviklingen af Fedoras funktionalitet. DTIC's bidrag på DEFF's vegne har været Fedoras søgefunktion, GSearch, der indgår i den centrale distribution og anvendes verden rundt.

Med etableringen af DuraSpace som fælles organisation for såvel Fedora som det andet store repository system, Dspace, er DEFF's vision om en fælles generisk software løsning for repositories ved at blive opfyldt. DuraSpace har nemlig besluttet at reimplementere Dspace som et brugerinterface / frontend oven på Fedora. Herved gør Dspace følgeskab med et stort antal lignende repository frontends som eSciDoc, Hydra, Islandora, Fez, Statsbibliotekets DOMS, australske ARROW, etc.

DEFFs hidtidige støtte til Fedora-arbejdet har været meget vellykket, bl.a. med følgende resultater: Dansk ledelse af udviklingen af søgefaciliteter i Fedora; Dansk analyse af og bidrag til Fedoras bevaringsfaciliteter; Dansk medlemskab af Fedora Advisory Board; Dansk lederskab i organiseringen af det europæiske user community; Dansk deltagelse i en vejledende arkitekturgruppe i marts 2007; Ledelse og afholdelse af workshops på European Conference on Digital Libraries [5][6]; Præsentationer ved Open Repositories konferencer [7][8]; og generelt Dansk tilstedeværelse i Fedoras community.

Det er således af såvel national som international betydning, at GSearch opdateres, således at den lever op til moderne forventninger og integrerer optimalt med den nyeste Fedora version.

## 6.2 Mål

Formålet opnås gennem opfyldelse af en række delmål:

- Opdatering til de nyeste komponent-biblioteker.
- Videreudvikling af funktionaliteten på basis af opsamlede og prioriterede brugerønsker.
- Udvikling af reference-konfigurationer til støtte for implementeringen af nye Fedora-applikationer, inklusive en tutorial.
- Forbedring af performance, baseret på omfattende stress-tests.
- Formidling og support af GSearch via Fedora mailinglisten.

4

## 6.3 Målgruppe

System- og service-udviklere ved forskningsbiblioteker mv.

## 6.4 Strategi/metode

Udviklingen af GSearch softwaren benytter samme strategi og metode som udviklingen af Fedora og andet open source software, herunder at man hyppigt opdaterer de komponent-biblioteker, som man benytter fra andre open source projekter. Man benytter kode repositories som svn eller nu git, og man benytter løbende udvikling af test suiteer til sikring af kvaliteten. Desuden er feedback fra brugersamfundet afgørende.

Den vigtigste open source komponent i GSearch er Lucene [9], som er en indekserings- og søgemaskine fra Apache organisationen. Den næstvigtigste open source komponent i GSearch er Solr [10], også fra Apache, som er en implementering af en server for Lucene. GSearch brugere kan vælge at bruge Lucene alene eller sammen med Solr. Mange forskningsbiblioteker, som bruger GSearch, foretrækker Lucene-med-Solr løsningen. Lucene og Solr følges nu i versionsudviklingen, og nyeste version er nu 3.5.0.

## 6.5 Succeskriterier/resultater

At en opdateret og forbedret GSearch indgår i Fedoras globale standard-distribution.

## 7 Præsentation af projektet

### 7.1 Beskrivelse af projektføreløbet

Projektet indledtes med forberedelse til og deltagelse i Open Repositories Conference 2011 [14], som blev afholdt 6.-11. juni 2011 i Austin, Texas. Konferencen og specielt Fedora User Group dagene er den vigtigste årlige begivenhed for Fedora-samfundet, og dermed også for brugerne af GSearch.

Projektet havde taletid på grundlag af et Proposal indsendt i februar 2011, og indlægget havde titlen "FedoraGenericSearch – Status and Future Developments".

Præsentationen [15] beskrev det planlagte indhold af to nye Releases af GSearch, hvoraf den første fokuserede på Out-of-the-box brugeren, som ønsker på nemmest mulige måde at gøre objekterne i et Fedora repository søgbare. Den anden Release fokuserede på det avancerede bruger-team, som ønsker fuld udnyttelse af søgemaskinen, hvilket kræver overblik og indsigt i langt flere detaljer.

Præsentationen indeholder også den bedste introduktion til GSearch, der findes.

Den egentlige realisering af planen begyndte, som normalt i Fedora-projektet, med oprettelse af Issues (også kaldet Tickets), se listen af Issues [16]. Under udviklingsarbejdet benyttes git [17] og github [18], hvor den udviklede kode lægges op, og der krydsrefereres mellem koden og Issues.

Første nye Release blev GSearch version 2.3 i september 2011. Anden og sidste nye Release i projektet blev GSearch version 2.4 i december 2011. Det fremgår af Issue-listen, hvad der er blevet realiseret hvornår. Releases er open source og kan downloades fra Fedora Commons sitet [19] [20].

Dokumentationen er inkluderet i Releases i form af en omfattende web-side, hvortil kommer en række konfigurationsfiler, som indeholder forklaringer til brugerens muligheder i mange detaljer.

Releases er annonceret på Fedora Commons mail-listerne [21] [22], hvor mange diskussioner, spørgsmål og svar kan ses.

### 7.2 Beskrivelse og evaluering af projektets resultater

Projektets delmål er realiseret gennem udførelse af Issues således:

#### 7.2.1 Opdatering til de nyeste komponent-biblioteker

De nyeste komponenter er Fedora Commons 3.5, Lucene 3.5.0, Solr 3.5.0, Zebra 2.0.46 og PDFBox 1.6. Dette er opnået med FCREPO-980, FCREPO-981, FCREPO-982, FCREPO-983 og FCREPO-1005.

#### 7.2.2 Videreudvikling af funktionaliteten på basis af opsamlede og prioriterede brugerønsker

Følgende Issues bidrager til dette:

- FCREPO-979 Simplified configuration for out-of-the-box users
- FCREPO-992 Selection between xslt processors, xalan or saxon

- FCREPO-996 Add setAllowLeadingWildcard to configuration of the Lucene plugin
- FCREPO-1006 Useful end-user search page generation from indexing stylesheet
- FCREPO-1008 Filtering of search results by access constraints
- FCREPO-1009 Interaction with the Resource Index
- FCREPO-1010 Use Apache Tika for extraction
- FCREPO-1018 Management of GSearch configurations in Fedora objects
- FCREPO-1019 Exploration of complex GSearch use cases

### **7.2.3 *Udvikling af reference-konfigurationer til støtte for implementeringen af nye Fedora-applikationer, inklusive en tutorial***

FCREPO-979 samt dokumentations-Issues FCREPO-972 og FCREPO-1028 understøtter disse delmål.

Præsentationen ved OR2011 [15], sammen med dokumentationssiden og konfigurationsfilerne i downloadet [20] udfylder behovet for en tutorial.

### **7.2.4 *Forbedring af performance, baseret på omfattende stress-tests***

FCREPO-1007 indeholdt arbejdet med performance tests, som blev opsamlet og kommenteret i en rapport for sig, som er inkluderet GSearch 2.4, se [23]

### **7.2.5 *Formidling og support af GSearch via Fedora mail-listerne***

Releases er annonceret på Fedora Commons mail-listerne [21] [22], se Bilag 3.

Mange emner har været bragt op og er blevet besvaret, fx:

- Nelson Hart [nhart@upei.ca](mailto:nhart@upei.ca) : Setting up Fedora Repo on a different ip then Fedora Gsearch
- Serhiy Polyakov [sp0055@gmail.com](mailto:sp0055@gmail.com) : FedoraGsearch 2.3 index update
- Phil Cryer [phil.cryer@mobot.org](mailto:phil.cryer@mobot.org) : Ingested items not showing up in Solr
- Caleb Derven [caleb.derven@ucd.ie](mailto:caleb.derven@ucd.ie) : Gsearch/ solr indexing issue
- Frank Feng [frank.feng@york.ac.uk](mailto:frank.feng@york.ac.uk) : More configurations with Fedora Gsearch while changing to external ActiveMQ?
- Swithun Crowe [cs2@st-andrews.ac.uk](mailto:cs2@st-andrews.ac.uk) : GSearch and FeSL/JAAS
- Scott Hammel [prov356@g.clemson.edu](mailto:prov356@g.clemson.edu) : note on GSearch and MaxPermSize
- Robert Rice [robert.rice@yale.edu](mailto:robert.rice@yale.edu) : Gsearch v2.3
- Stuart Chalk [schalk@unf.edu](mailto:schalk@unf.edu) : Indexing PDF files
- Dave Raskin [dave.raskin@rimage.com](mailto:dave.raskin@rimage.com) : indexing managed data stream
- Serhiy Polyakov [sp0055@gmail.com](mailto:sp0055@gmail.com) : Fedora GSearch and command line Xalan processing with foxmlToSolr.xsl
- Matthias Hahn [Matthias.Hahn@fiz-Karlsruhe.de](mailto:Matthias.Hahn@fiz-Karlsruhe.de) : Gsearch and Fedora 3.5 with Hadoop
- Scott Hammel [scott@clemson.edu](mailto:scott@clemson.edu) : next release of GSearch
- Arash Samadi [asamadi@sub.uni-goettingen.de](mailto:asamadi@sub.uni-goettingen.de) : Fedora 3.5-SNAPSHOT and GSearch 2.2
- Menk, Robert - 1150 - MITLL [bmenk@ll.mit.edu](mailto:bmenk@ll.mit.edu) : gSearch installation questions for Fedora 3.4.2
- [Soja@soja.ir](mailto:Soja@soja.ir) : A question about Gsearch service
- Federica Bertozzi [f.bertozzi@inera.it](mailto:f.bertozzi@inera.it) : GSearch: how to remove from index deleted object
- Swithun Crowe [cs2@st-andrews.ac.uk](mailto:cs2@st-andrews.ac.uk) : problems with GSearch + Solr

- Foudil BRÉTEL [Foudil.Bretel@inria.fr](mailto:Foudil.Bretel@inria.fr) : combining GSearch + RISearch ?
- Matteo Bertazzo [m.bertazzo@cineca.it](mailto:m.bertazzo@cineca.it) : gSearch and GSA integration
- Nilani Ganeshwaran [Nilani.Ganeshwaran@manchester.ac.uk](mailto:Nilani.Ganeshwaran@manchester.ac.uk) : Deleting one index document
- Wolff, Robert [Rob.Wolff@unh.edu](mailto:Rob.Wolff@unh.edu) : Gsearch/Solr indexing problem
- Benjamin Ryan [B.Ryan@leeds.ac.uk](mailto:B.Ryan@leeds.ac.uk) : Indexing updates with 3.4 and SOLR
- Christopher Curry [ccurry@amphilsoc.org](mailto:c Curry@amphilsoc.org) : Configuring GSearch to use Solr
- Matteo Boschini [matteo.boschini@gmail.com](mailto:matteo.boschini@gmail.com) : dummy question on gsearch
- Alistair Young [alistair.young@uhi.ac.uk](mailto:alistair.young@uhi.ac.uk) : gsearch 2.2 wsdl breaks with axis2
- Scott Hammel [scott@clemson.edu](mailto:scott@clemson.edu) : metadata theory / practice

## 8 Konklusion

Projektet har været meget vellykket. Det har produceret to nye Releases af GSearch som planlagt og præsenteret for Fedora samfundet. Overordnede og detaljerede forklaringer og vejledninger er indeholdt i OR2011-præsentationen og i downloadet. Fortløbende kommunikation med særligt interesserede brugere har medført flere nye features. Der er lagt op til yderligere kommunikation med brugerne på maillisten og direkte, og der vil efterfølgende blive fulgt op på eventuelle problemer og ønsker til nye features. Et paper om projektet vil blive indsendt til Open Repositories 2012 konferencen i Edinburgh i juli 2012.

Med afslutningen af dette projekt afrundes således DEFF's engagement i udviklingen af søgefaciliteter til Fedora. Der skønnes ikke at være behov for yderligere projektstøtte fra DEFF's side. Da flere DEFF-biblioteker (ikke mindst KB, Statsbiblioteket og DTIC) fortsat er tunge brugere af Fedora, vil samspillet med Fedoras Core Team og de øvrige teknisk tunge brugere givetvis fortsætte.

Afslutningsvis, et citat fra Fedora maillisten, efteråret 2011:

- Scott Hammel [scott@clemson.edu](mailto:scott@clemson.edu) : metadata theory / practice  
(Quote: "Yeah ... a vote of confidence for the work Gert and his team are doing: GSearch takes a lot of the headache out of indexing any XML datastream (or combination of them) on your objects into a powerful search index (and with Solr you get some geospatial index/query helpers).")

## 9 Links

- [1] <https://wiki.duraspace.org/display/FCSVCS/Generic+Search+Service+2.4> - Hjemmeside for GSearch, hvor princippet er forklaret i den indeholdte web-side, vedlagt som Bilag 2
- [2] <http://www.deff.dk/showfile.aspx?IdGuid=%7B4F681F6B-397D-467D-9744-BA669A8B9B7B%7D> - Afrapportering for DEFF projektet: Fedora som generisk repository arkitektur, 2005
- [3] <http://www.deff.dk/showfile.aspx?IdGuid=%7B0BCF2A99-F31E-4B8C-9A3B-EADEA7AABD7C%7D> - Dansk deltagelse i den videre udvikling af Fedora, DEFF afslutningsrapport, Oktober 2009
- [4] <http://fedora-commons.org/> - Hjemmeside for Fedora Commons

- [5] <http://www.lib.uoa.gr/dorsdl/> - 1<sup>st</sup> European Workshop on the use of Digital Object Repository Systems in Digital Libraries (DORS DL)
- [6] <http://dorsdl2.cvt.dk/> - 2<sup>nd</sup> European Workshop on the Use of Digital Object Repository Systems in Digital Libraries (DORS DL2)
- [7] <http://pubs.or08.ecs.soton.ac.uk/104/> - Præsentation ved Open Repositories Conference 2008
- [8] <https://or09.library.gatech.edu/fedora.php> - Præsentationer ved Open Repositories Conference 2009
- [9] <http://lucene.apache.org/> - Hjemmeside for Lucene
- [10] <http://lucene.apache.org/solr/> - Hjemmeside for Solr
- [11] <https://wiki.duraspace.org/display/FCREPO/Fedora+Contributors>
- [12] <http://www.kemibrug.dk/> - Kemiportalen, DTIC anvendelse af Fedora og GSearch
- [13] <https://conferences.dtu.dk/Proceedings> - Conference Proceedings, DTIC anvendelse af Fedora og GSearch
- [14] <https://conferences.tdl.org/or/OR2011/OR2011main> - Open Repositories Conference 2011
- [15] <https://conferences.tdl.org/or/OR2011/OR2011main/paper/view/416/127> - Præsentation på OR2011, vedlagt som Bilag 1
- [16] <https://jira.duraspace.org/secure/IssueNavigator.jspx?mode=hide&requestId=10311> - Liste af GSearch issues
- [17] <http://git-scm.com/> - git is a free & open source, distributed version control system
- [18] <https://github.com/fcrepo/gsearch> - git repository for GSearch
- [19] <https://wiki.duraspace.org/display/FCSVCS/Generic+Search+Service+2.3> - Download af GSearch 2.3
- [20] <https://wiki.duraspace.org/display/FCSVCS/Generic+Search+Service+2.4> - Download af GSearch 2.4
- [21] [fedora-commons-users@lists.sourceforge.net](mailto:fedora-commons-users@lists.sourceforge.net) - Support and info exchange list for Fedora users
- [22] [fedora-commons-developers@lists.sourceforge.net](mailto:fedora-commons-developers@lists.sourceforge.net) - Fedora developers mail list
- [23] <https://github.com/fcrepo/gsearch/tree/master/FedoraGenericSearch/src/performance> - rapport over performance tests

## 10 Bilag

1. Præsentation på Open Repositories Conference 2011 [15]
2. Dokumentation for Fedora GSearch version 2.4 [20]
3. Mail til Fedora-user-listen om release af GSearch 2.4