Fedora, DSpace og de andre

Kort gennemgang af Open Source arkivsoftware - bl.a. til brug for "Institutional Repositories"

Af Alfred Heller

I de senere år tales der om begrebet "Institutional Repositories" (IR), som skal give forskningsinstitutterne mulighed for at arkivere deres forskningspubliceringer, gøre dem tilgængelig for et publikum, samt hjælpe med bevaringsopgaven. Der er udviklet computerprogrammer, der klassificeres som IR, selv om ikke alle er begrænset til anvendelsen til dette formål.

Alfred Heller
Ph.d. og civilingeniør
Systemarkitekt og Specialkonsulent
Center for Videnteknologi (CVT)
Danmarks Tekniske Videncenter (DTV)
ajh@cvt.dk

Introduktion

Før vi er klar til at se på den software, der i almenhed anses som "Institutional Repository" citeres her en af de definitioner, der kunne danne basis for udtrykket:

"A university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution." (www.dspace.org)

Jeg mener, at denne definition fanger de centrale egenskaber, der danner de specifikationskrav, vi anvender for udvikling af de computersystemer, som ønskes af universiteterne. Der kunne dog tilføjes ønsket om integration af IR-komponenten i de eksisterende infrastrukturer, dvs. andre computersystemer.

Fedora og DSpace har det til fælles, at de er udviklet i programmeringssproget Java under Open Source paradigma, at de stammer fra berømte universiteter beliggende i USA, samt at de er finansieret af samme fond, The Andrew W. Mellon Foundation. Ellers er det svært at finde en fællesnævner. På trods af dette forhold, bliver de ofte nævnt i flæng og oftest som (Digital) "Institutional Repositories". Et af formålene med den nærværende artikel er, at afløse denne sammenblanding med oplysning. Herudover er formålet naturligvis også, at give en gennemgang af værktøjerne og de erfaringer, der foreligger hos forfatteren og dennes organisation.

DSpace er en færdig softwarepakke, som berettiget kan klassificeres som "Institutional Repository" (IR). Fedora derimod er ikke en færdig applikation, og endnu mindre et færdigt IR-system. Det er mere et værktøj/platform til håndtering af alle mulige slags digitale objekter og deres indbyrdes forhold. Det kan dog forventes, at der indenfor et års tid vil findes applikationer til bl.a. IR-formål, som vil være bygget på Fedora. Mere herom under afsnittet om Fedora.

Der findes andre systemer, som kan klassificeres som IR. Det nok mest anvendte er vel E-prints som danner basis for arXiv[5], verdens nok største preprint service indenfor emner som f.eks. fysik, matematik og computer forskning. Endvidere kan det nævnes, at der i New Zealand er udviklet et alternativ, som hedder Greenstone[4]. Programmet henvender sig til brugere med mindre sofistikerede behov og er i den nuværende version kun anvendelig for simple web-kollektioner. I en planlagt version vil dette dog ændre sig, hvis det kommer dertil.

På DTV har vi forsøgt os med Fedora og DSpace, men kun set på en ældre version af Greenstone. Konklusionen fremføres nedenfor.

E-print

E-prints[1] er udviklet på University of Southampton, UK. På websiden fremgår det, at der findes 161 arkiver med ca. 86.000 records. Programmet bliver altså anvendt i større stil, hvilket ikke helt kan siges om nogle af de andre softwarepakker.

E-prints er udviklet i programmeringssproget "Perl". Systemet har nok sammen med Greenstone[4] de mindste systemkrav. Installationen kræver færdigheder i kompilering m.m. Det gælder dog også de andre nævnte systemer, men det er en overkommelig udfordring for de fleste. Programmet kan i høj grad konfigureres såvel i forhold til funktion, som i forhold til udseende. Konfigureringen skal være på plads før ibrugtagning, da mange egenskaber ikke kan ændres, når systemet er i drift. (Det er ikke anderledes for DSpace). E-prints tilbyder simple web-brugergrænseflader, der hjælper med administration og anvendelse af systemet.

Da undertegnede ikke selv har installeret e-print, så overlader jeg læserne til egne undersøgelser gennem linkene givet nedenfor, bl.a. gennem et link til beskrivelse af implementering af et e-print system.

DSpace

Dspace[2] er udviklet i samarbejde mellem MIT og Hewlett-Packart Company, og er født som Institutional Repository, som det er beskrevet på deres hjemmeside, hvorfra citatet i indledningen er hentet.

Installationen af DSpace er rimelig lige til, hvis man har erfaringer i Java programmer. Der skal lidt konfigurering til og der kan anvendes en del "tricks" til effektivisering af denne proces.

Har man installeret softwaret, er næste trin at tilpasse programmet, så det passer til ens egen anvendelse. Programmet definerer en "enhed" (item) på en måde, hvor en given metadata beskriver en eller flere digitale filer (pit streams). Samlinger af digitale objekter kaldes "Collections", som hver har sin egen brugergrænseflade (side). Dermed kan man browse og søge i en enkelt collection eller på tværs af disse. Til brug for specielle forhold kan man etablere "virtuelle collections", som ser ud til at være samlinger, der dog kun eksisterer som en slags tværgående søgninger gennem systemets andre samlinger.

Brugere kan opdeles i 2 grupper - brugere, der aktivt lægger data ind i systemet (submitters) og brugere, der ser på dataene (end users). Submitters organiseres i "Interessegrupper",

DSpace er en færdig softwarepakke, som berettiget kan klassificeres som ”Institutional Repository” (IR). Fedora derimod er ikke en færdig applikation, og endnu mindre et færdigt IR-system.

såkaldte “Communities”, evt. opbygget med hierarkiske sammenhænge som selvstyrende grupper, der hver kan have sin måde at deponere og kvalitetssikre sine data på. Som supplement hertil, tilbyder programmet “virtuelle” samlinger, som baseres på objekter, der ligger på tværs af de organisatoriske samlinger. Til alle de nævnte procedurer, er der udviklet ret nemme og brugervenlig grænseflader.

Når man har opbygget sine grupper og samlinger, skal rettigheder for brugerne i forhold til samlinger og interessegrupper afklares. Her tilbyder DSpace et brugervenligt web interface, der dog straks bliver for omstændeligt at bruge, når man har mange brugere. Nedenfor nævnes arbejde, der er lavet for at løse denne opgave mere effektivt.

Er modellerne for samlinger og grupper introduceret, er det meget vigtigt at fremhæve, at den centrale, og nok mest krævende opgave for opbygning af et IR-system er, at kunne overføre de konkrete forhold for et nyt IR til de modeller, der ligger for det givne software. Man kan opbygge sine grupper således, at de gengiver de organisatoriske forhold på den givne institution, f.eks. fakulteter, institutter og forskningsgrupper. Man kan også klassificere sine data efter type, f.eks. rapporter, fuldtekst-artikler, billeder m.m. Mulighederne er mange og kræver et grundigt forarbejde. Det tager lidt tid at implementere modellen i DSpace, men svært er det ikke. Når alt dette er på plads, er lidt ”smarten up” af brugergrænsefladerne oplagt, så ens eget depot ikke ser ud, som alle de andres. Nu er systemet klart til slutbrugerne, som har adgang til arkivet gennem den brugergrænseflade, som er opbygget for det givne system.

Adgangsberettigede brugere kan have en eller flere roller i forhold til en eller flere samlinger/grupper. Man kan være administrator, editor, peer reviewer eller submitter, samt kombinationer heraf. En submitter kan, efter login, vælge mellem de samlinger, han/hun har rettigheder til. Hvis man vil aflevere en publikation eller lignende, vil man blive bedt om at udfylde en række skemaer, der tilsammen udgør et workflow, som afsluttes med accept af de licensbetingelser, der knytter sig til deponering af filer, hvorefter de uploades. De deponerede objekter vil indgå i et workflow afhængigt af de aftaler, en given gruppe eller samling har aftalt. Har man aftalt, at der forekommer et peer-review, så vil dokumentet indgå i den “kø” af dokumenter, som står til review. Alle udvalgte peers vil blive tilbudt jobbet gennem brugergrænsefladen, men også gennem en mail-alert. ”Først-til-mølle princippet” gælder her - er dokumentet accepteret, vil det blive synligt for offentligheden.

Som nævnt, skal man tildele en bruger rettigheder for hver gruppe og hver samling. Dette er en omfattende opgave, selv om man har en brugergrænseflade hertil. For et universitet er det ikke et realistisk scenario. Derfor har DTV, i samarbejde med Nordija A/S, udviklet to alternative måder at styre brugerrettigheder på. Den ene anvender en lokal CAS single sign-on, den anden giver mulighed for at overføre rettigheder fra en database eller lignende til DSpace databasen. Den sidstnævnte anvendes på RUC, mens førstnævnte er udviklet til DTU.

DSpace gør det meget let at udføre de opgaver, programmet er skrevet til. Det er dog straks en anden sag, hvis man vil ændre på tingene. Det begynder med sproget, som er engelsk og ret besværlig at ændre til et andet sprog. Flere sprog på en gang er ikke muligt. Man kommer hurtigt til det punkt, hvor man skal tage fat i egentlig programmering. For enkelte applikationer er der udviklet udvidelser, som løser mere omfattende ændringer. “Tapir” er en af de mere kendte eksempler herpå, hvor man har tilpasset DSpace til at være et Thesis-håndteringsværktøj med vejledningsprocedure m.m.

Alt i alt er DSpace et godt produkt, hvis man bruger det til det, det er tænkt til. Det kan dog næppe bruges til andre formål, hvis ikke en genprogrammering i modulær form gennemføres, hvilket vi har ventet på i et år eller to. Der arbejdes på en ny version, som skulle kunne løse denne begrænsning. Da den ikke er i udsigt efter 1-2 år, er det usikkert, hvordan det går med dette projekt. Der bliver endda talt om, at man bruger Fedora som basis for en mere modulær version af DSpace. Lad os se.

Anvendelseseksempler i Denmark

På RUB anvendes softwaret til opsamling af studenterpublikationer m.m. Der er opbygget en installation, hvor hver organisatorisk enhed har sin egne “Communities” og herunder er der opbygget kollektioner efter et fast skema, f.eks. dokumenttype. Hele afleveringsproceduren er et kapitel, der ikke gennemgås her, men som nok er det emne, der kræver flest ressourcer af projektledelsen - langt mere end at installere softwaren og drive servicen.

På DTU har DTV installeret en testversion af DSpace for at undersøge, om man kan bruge DSpace som arkivkomponent til forskningsdatabasen. Det viste sig, at samarbejdet med udviklerne i USA var vanskeligt og at de fordrede ressourcer til selv at udvikle løsningen ikke var til stede. Herefter er softwaren fravalgt som mulig komponent til den overordnede infrastruktur.

Fedora er mest af alt en database, der giver flere muligheder for håndtering af digitale objekter på forskellige måder, endda måder som designerne af Fedora ikke kunne vide noget om.

Som det nok fremgår af teksten, findes der for tiden ikke et egentligt software, der lever op til de krav, man stiller til et fleksibelt, let anvendeligt og let installerbart software, der kan anvendes til de mange arkivopgaver, der ligger foran os.

Fedora

Fedora[3] er udviklet i samarbejde mellem Cornell University og University of Virginia Library, og er snarere et værktøj til udvikling af applikationer. Programmet tilbyder grundlæggende muligheder for deponering og håndtering af digitale objekter og deres indbyrdes forhold. Når der findes mindre applikationslignende komponenter, er de mest tænkt som demonstratorer og minimalistiske værktøjer til udviklere. Altså den diamentrale modsætning til DSpace – det er ikke færdig redigeret til anvendelse. Herefter kunne gennemgangen af programmet i forhold til IR stoppe, dog er det den generiske egenskab, som gør programmet til det nok mest relevante software-produkt for digitale biblioteker og lignende. For at afrunde relationen af Fedora til IR, så forventes det, at man kan hente færdige applikationer indenfor det kommende års tid bl.a. til en IR-applikation, der udnytter Fedora som "motor".

Fedora er mest af alt en database, der giver flere muligheder for håndtering af digitale objekter på forskellige måder, endda måder som designerne af Fedora ikke kunne vide noget om. Dette er muligt gennem anvendelse af "Web Services" (WS). WS er en metode, hvorigennem systemer kan "tale sammen" og arbejde sammen bl.a. ved at anvende XML som fælles datarepræsentation. Fedora bruger WS til intern kommunikation mellem systemkomponenterne. Hermed er det muligt at erstatte dele, der kommer med Fedora med egne komponenter, fx erstatte søgedelen eller lade ens eget Content Management System vise resultater fra de objekter, der ligger i Fedora. Dette opnås ganske simpelt ved at pege på URL-links.

Med Fedora installationen følger en række værktøjer, som gør det muligt at fodre den med digitale objekter og få dem vist på en standardmåde i en browser. Man kunne fx lave sine metadata i en fil for sig og tilknytte et digitalt billede. Som standard vil man kunne se metadata som XML, som Dublin Core og se billeder i den originale form. Det er ret nemt at få lavet thumbnails og web-optimerede billeder. Man kan søge sine objekter igennem, dog i begrænset omfang. Hermed stopper standardudgaven og det er op til andre at anvende systemet til de formål, de har. Eksempler er der mange af.

Hamletworks.org er et site, hvor sprogforskere studerer Hamlets værker. En interessant egenskab er, at web-applikationen anvender et plug-in (Djvu), som gør det muligt at zoome ind på en tekstscanning i en meget høj scanningsopløsning og derved studere teksten med en uset nuancering, som ellers ville kræve både mikroskop og originalteksterne.

The Encyclopedia of Chicago er en mere populær anvendelse af Fedora. Her kan man finde en hel masse digitale objekter som indscannede tekster i forbindelse med digitale kort. Med disse kan man arbejde frem og tilbage mellem tekstuelle beskrivelser og kort af hændelser, der foregik i Chicago indenfor de seneste århundreder. Fedora er i denne applikation udnyttet til at håndtere hele processen fra lagring til fremvisning af den endelige webpublikation.

En tænkt anvendelse af Fedora har til formål at facilitere arkiver med et system, der kan understøtte præservering/langtidsbevaring. Her kan man forestille sig, at man udvikler metoder til formatering af givne objekttyper til mere nutidige versioner. Fedora vil benytte denne metode for alle objekter af en given type, lade objektet konvertere og lagre den nye version ved siden af den gamle version, samt opdatere de nødvendige metadata, så de passer til den nye version.

Som man nok kan fornemme, så har Fedora alverdens anvendelsesmuligheder og det er en af de problemstillinger, man skal gøre sig klart. Der findes endnu ikke for mange gennemarbejdede anvendelsesmuligheder. Der arbejdes på at opsamle de få erfaringer på Fedoras hjemmesider, man indtil nu har gjort. Indenfor et års tid vil vi se de første færdige applikationer af Fedora "out of the box". Man vil kunne finde DSpace-lignende IR-applikationer, applikationer til arkivering og vedligeholdelse, helt almindelige Content Management Systemer, samt mange ret sofistikerede applikationer, der vil blive anvendt af de få til nogle helt specielle formål.

Anvendelseseksempler i Denmark

DTV har demonstreret anvendelse af Fedora som et e-print-arkiv. DTV er i gang med at integrere indekseringsværktøjet Zebra i Fedora-komplekset. Hermed vil søgemulighederne for programmet øges betragteligt.

DTV arbejder bl.a. på at få etableret et sådant ved at sammenbygge MetaToo-softwaret, der bl.a. anvendes i "Den Danske Forskningsdatabase" og som er et helt fleksibelt, konfigurerbart værktøj til indsamling af metadata og dertil hørende digitale objekter. En anden komponent vil være at erstatte SQL-søgemaskinen fra Fedora med Zebra fra IndexData, hvorefter Fedora vil kunne supportere SRU/SRW-søgeprotokollerne m.m.

Det kan nævnes, at der blev afholdt en europæisk Fedora-user-konference den 28. september i København (http://fedora.cvt.dk).

Afslutning

Som det nok fremgår af teksten, findes der for tiden ikke et egentligt software, der lever op til de krav, man stiller til et fleksibelt, let anvendeligt og let installerbart software, der kan anvendes til de mange arkivopgaver, der ligger foran os. DSpace kan bruges til lige præcis de opgaver, den er lavet til, samt de få opgaver, der er udviklet udvidelser til. Fedora er ikke tilpasset til konkrete anvendelser og vi må vente på dem. Greenstone mangler vi erfaringer med, men den nærværende version er kun anvendelig for online samlinger. Alligevel er fremtiden lovende, da de sidste års erfaringer har givet et fingerpeg om, hvordan de kommende systemer skal opbygges og fungere. Vi kan håbe, at disse erfaringer kan omsættes og bringe os de længe ventede hjælpeværktøjer vi har brug for.

Referencer

[1] E-print
www.eprints.org, http://software.eprints.org og
www.ariadne.ac.uk/issue31/eprint-archives/
[2] DSpace
www.dspace.org
[3] Fedora
www.fedora.info
[4] Greenstone
www.greenstone.org
[5] ArXiv
www.arxiv.org
[6] Fedorakonference
http://fedora.cvt.dk