Projekti o digitalnim dokumentima

Dragan Ivanović
dragan.ivanovic@uns.ac.rs

Katedra za informatiku, Fakultet tehničkih nauka, Novi Sad

2015.

Sistemi i naučni sadržaji

- Informacioni sistem naučno-istraživačke delatnosti
  - projekti
  - institucije
  - istraživači
  - publikovani rezultati (naučni sadržaji)
  - oprema
  - ...
- CRIS - CERIF
- Institucionalni repozitorijumi, bibliotečki informacioni sistemi, NDLTD, DART-Europe, ...

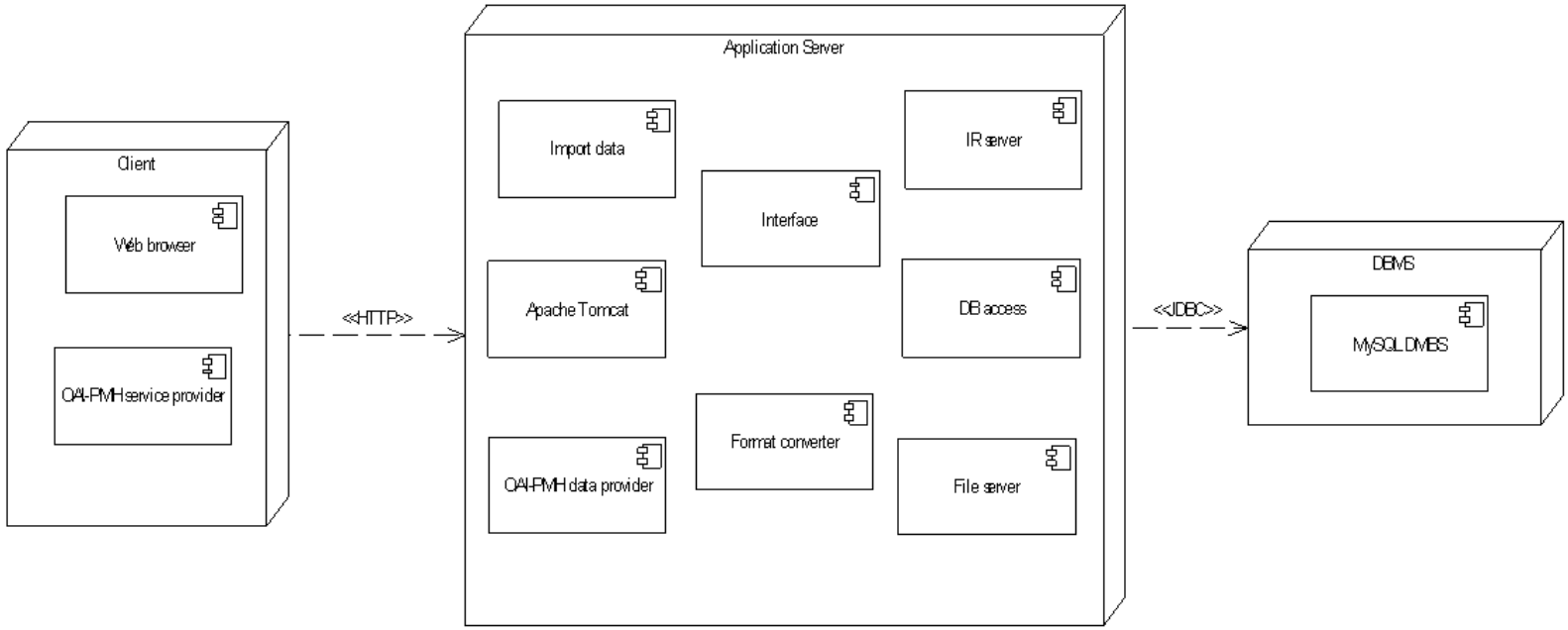
CERIF kompatibilni model podataka

- Naučni sadržaji se skladište u MARC 21 formatu
- Razmena podataka sa
  - CRIS sistemima
  - bibliotečkim sistemima zasnovanim na MARC 21 formatu
  - institucionalnim repozitorijumima zasnovanim na Dublin Core formatu
  - Članovima NDLTD mreže
- Model proširen potrebnim entitetima za vrednovanje po našem pravilniku

Opšti podaci

- http://www.cris.uns.ac.rs
- Koristi ga PMF-Novi Sad
- Univerzitetski servisi
- I dalje se razvija
- Interoperabilan sa svetskim sistemima, a zadovoljava i domaće potrebe

Arhitektura



Naši naučni sadržaji

- Radovi na konferencijama
- Radovi u časopisima
  - Jedan rad u vrhunskom međunarodnom časopisu
  - Dva rada u istaknutim međunarodnim časopisima
  - Pet radova u međunarodnim časopisima
  - Dva rada u nacionalnim časopisima
  - Dva poslata rada u međunarodnim časopisima
- Monografija - Andrejević
- Doktorati
  - Jedan odbranjen
  - Jedan predat na uvid javnosti
  - Tri u izradi

Zapisnici sa sednica

- Sednice se održavaju na nekoliko dana u APV
- Zapisnik sadrži broj sednice, datum održavanja, prisutne, odsutne, tačke dnevnog reda, ...
- Zapisnik je kreiran upotrebom MS Word alata
- Kako pretraživati i izveštavati po prethodno navedenim metapodacima

Predlog rešenja

- Formiranje zapisnika po usvojenom Word templejtu
- Ekstrakcija podataka iz zapisnika u Word dokumentu
- Prepis bibliotečki obrađenih Zapisnika u sistemu BISIS
- Zapisnike koji postoje samo u papirnatoj formi skenirati i bibliotečki obraditi u sistemu BISIS

Predlog rešenja

- Zapisnici se formiraju u obliku Word dokumenta i dostupni su samo u elektronskoj formi
- Zato se ovde prirodno nameće zahtev za definisanjem aktivnog dokumenta za formiranje zapisnika sa sednica Vlade APV
- Na osnovu ovako formiranog zapisnika automatski bi se generisala dva dokumenta (PDF format i XML verzija zapisnika)

Pretraga

- Osnovna i napredna pretraga
- Pronalaženje sednica sa svim rečima, sa identičnom frazom, sa najmanje jednom od reči i da ne sadrži određene reči
- Pretraživanje po predlagaču, predsedavajućim, dnevnom redu, potpisnicima, datumu *od do* i broju sednica u okviru saziva *od do*
- Korisnički interfejs podržava višejezičnost

Izveštavanje

- Broj sednica
- Broj tačaka i podtačaka dnevnog reda
- Ukupno materijala po vrstama
- Ukupno dostavljeno materijala
- Ukupno podnetih materijala po Sekreterijatima
- Prisutnost sednicama
- Predsedavanje sednicama

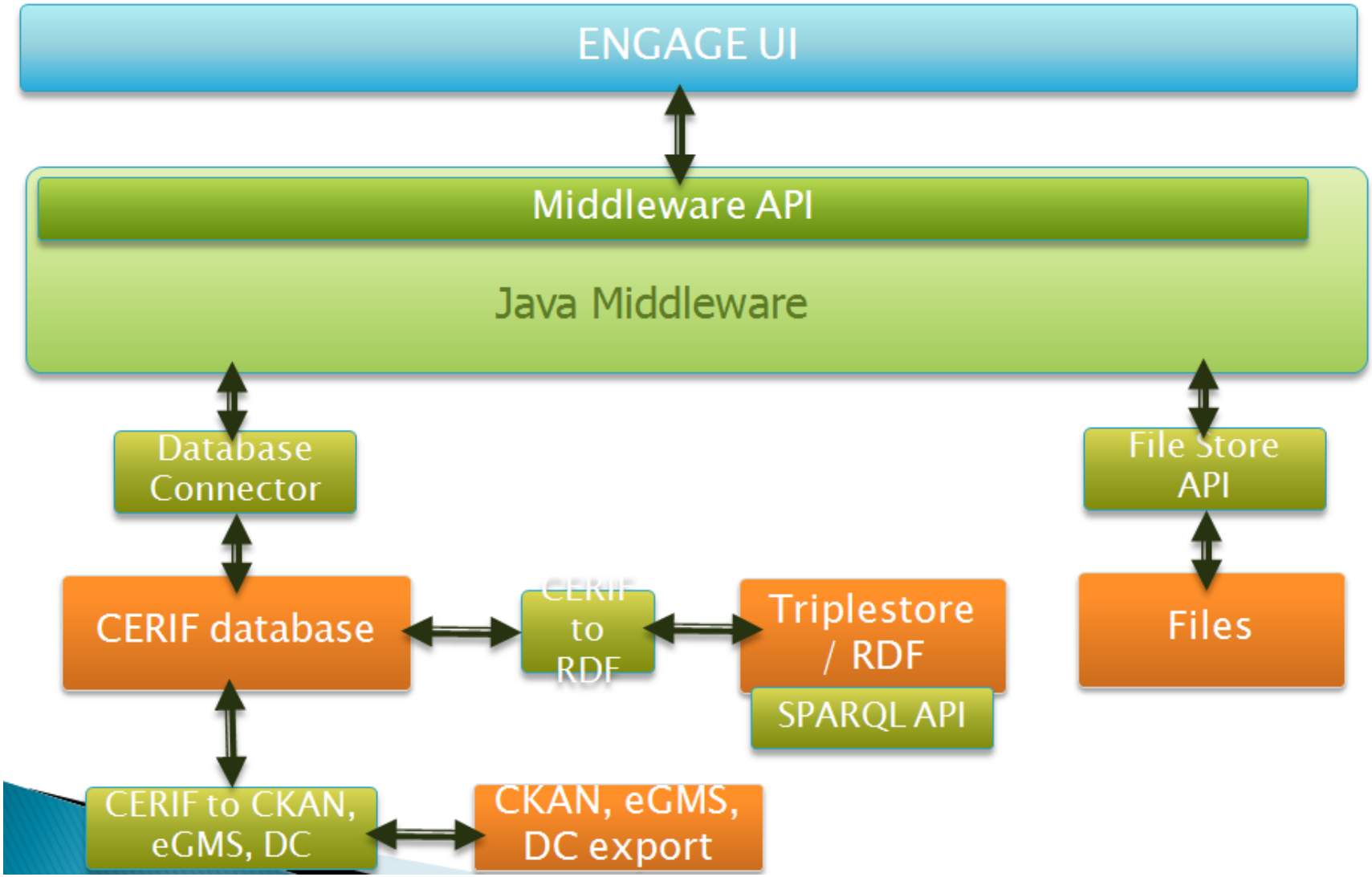
Status

- Instalirana je prva verzija u Izvršnom veću
- I dalje se razvija - zapisnici sastanaka radnih tela
- Monografija o ovom sistemu se piše

Opis projekta

- FP7 projekat – EU finansira
- Učesnici – Grčka, Engleska, Holandija, ...
- Jedan od učesnika euroCRIS
- Iz Novog Sada dva učesnika
- Cilj projekta povećati dostupnost metapodataka iz javnih sektora koji mogu biti od značaja za dalji razvoj nauke

Arhitektura



Engage Metadata Middleware 2.0

- Java
- MySQL
- Hibernate
- Spring
- Maven
- JPA Cerif
- REST API

Još interesantnih projekata

- Semantic keyword-based search on structured data sources (KEYSTONE)
  - COST akcija - EU finansira
  - Učesnici - preko 20 zemalja (Italija, Nemačka, Poljska, Španija, Srbija, itd.)
- HOLACloud
  - H2020 projekat – EU finansira
  - Jedan od učesnika euroCRIS
  - Cilj projekta povećati dostupnost rezultata projekata koji pripadaju oblasti računarskih nauka
- DOI Serbia
  - Dodela DOI brojeva digitalnim dokumentima
  - Uključuje i PhD disertacije - link

Hibernate i Lucene

- Implementacija repozitorijuma e-knjiga
- Indeksiranje se radi putem Hibernate anotacija
- Pretraživanje se radi pomoću Hibernate Search biblioteke
- Veb ili standalone aplikacija
- Uzeto!

Web search engine

- Periodično indeksiranje sadržaja odgovarajućeg Internet domena (npr. uns.ac.rs) preuzetog odgovarajućim crawler-om
- Implementacija veb aplikacije na kojoj se
  - zadaju upiti
  - dobijaju odgovori
- Za crawler se mogu koristiti gotova rešenja (npr. Apache Nutch)
- Uzeto i odbranjeno!

Search engine optimization

- Kreiranje veb sajta nekog softverskog proizvoda - proizvolja arhitektura i programski jezik
- Veb sajt treba da je optimizovan za pronalaženje sa Veb search engine-a (Google, Yahoo!) upotrebom ključnih reči koji pripadaju domenu softverskog proizvoda
- Analiza legalnih i spam tehnika
- Želja da se softverski proizvod lako pronalazi i da ga ljudi kupuju i koriste
- Uzeto i odbranjeno!

Snowball

- Programski jezik za rad sa stringovima, paterni određuju dalji tok izvršavanja
- Kreiranje i implementacija pravila za steming srpskih reči upotrebom Snowball-a
- Dobijeni stemmer je potrebno transformisati u Java kod (postoji gotov alat za ovu namenu)
- Dobijeni stemmer je potrebno i verifikovati i prikazati indikatore performansi
- Unakrsna validacija, postoji anotirani data set
- Uzeto!

Multi-lingual search

- Implementacija pretprocesora (analizatora) teksta koji omogućuje bilingualno pretraživanje repozitorijuma tekstualnih digitalnih dokumenata pisanih na Srpskom i Hrvatskom jeziku
- Ovo bilingualno pretraživanje je značajno jer postoji preko 10 miliona ljudi koji razume oba ova jezika
- Potrebno je kreirati jedinstveni stemer za oba jezika koji imaju slična morfološka pravila (postoji neka verzija koju je potrebno doraditi)
- Zatim koristiti rečnik koji sadrži reči koje su različite u ova dva jezika (postoji, ali i ovo treba doraditi)
- Analizator implementirati kao proširenje Lucene IR biblioteke
- Izvršiti verifikaciju kreiranog analizatora i prikazati rezultate.
- Unakrsna validacija, postoji anotirani data set

Long-term preservation - File Fixity

- Program za backup digitalnih dokumenata
- Inicijalno se vrši backup, odnosno dupliranje sadržaja i postavljanje u odgovarajući backup folder, kao i računanje fingerprint-a (jedinstvene niz bitova koji predstavlja dokument) nekim fingerprinting algoritmom
- Fingerprint se čuva do sledeće sinhronizacije radne verzije i backup-a
- Sinhronizacija je periodična
- Ako je radna verzija digitalnog dokumenta promenjena određenim programom (legalna izmena), računa se nova fingerprint vrednost i vrši backup
- Ako radna verzija nije menjana a došlo je vreme sinhronizacije, računa se fingerprint i ako je isti kao prethodna njegova vrednost sve je u redu, a ako nije ista onda je fajl oštećen
  - vršiti odgovarajuću notifikaciju
  - iz backup-a vratiti dokument

Lucene analyzer tester

- Razvoj aplikacije za testiranje Lucene Analyzer-a
- Kreiranje indeksa
  - indeksiranje sadržaja čiji su metapodaci u XML zapisima
  - indeksiranje digitalnih dokumenata koji se nalaze u odgovarajućoj folderskoj strukturi
  - importovanje već pripremljenih indeksa
- Postavljanje upita
  - importovati već pripremljene upite
  - korisnicima sistema dozvoliti da unose upite
- Rad sa rezultatima
  - korisnik sistema u listi odgovora označava relevantne i nerelevantne rezultate
  - sistem pamti odgovore korisnika
  - sistem računa preciznost, povrat, f-meru, kappa slaganje, itd.

Ostale teme

- OAI-PMH i Dublin core
- MG4J search engine
- JPlag, MOSS, Sherlock source code similarity detection tools
- Apache Tika content analysis toolkit
- Vaše ideje