Search & Develop

Marc Kemps-Snijders
Meertens Institute
Marc.kemps.snijders@meertens.knaw.nl

Utrecht
2011-08-25

Focus punten

- Ontwikkeling van een generieke oplossing voor zowel metadata als content search

  - Doelstellingen
    - Ontwikkeling van de oplossing
    - Opzetten van nationale CLARIN center struktuur
    - Ontwikkeling, validatie en stimulatie van CLARIN infrastruktuur
    - Verwerving van leidende positie binnen Europa

Metadata en data wat vragen wij?

Utrecht
2011-08-25

www.clarin.nl

Creatie van metadata documenten op basis van CMDI

Registratie van data categorieen

Registratie van CMDI componenten

\+ ISOcat + C Component Registry

Archivering van metadata en resources bij Clarin centra

The Language Archive

Meertens Instituut

DANS

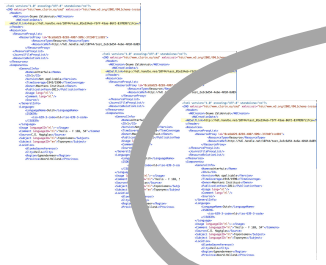
INL Instituut voor Nederlandse Lexicologie

huygens ing

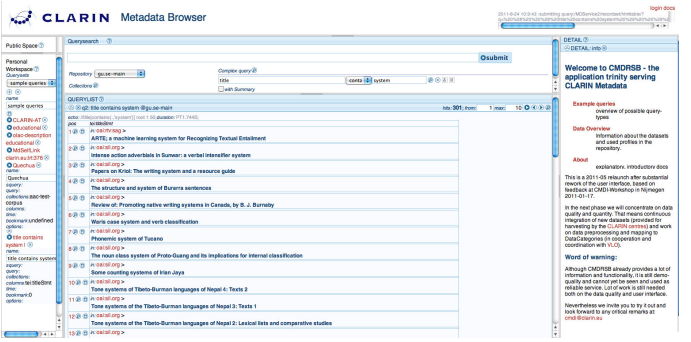
Metadata en data
wat doet Clarin ermee?

Utrecht
2011-08-25

www.clarin.nl



metadata harvesting
op basis van OAI-PMH



Afspraken

- Alle metadata binnen Clarin is CMDI gebaseerd.
- ISOcat en Relation registry worden gebruikt voor semantische interoperabiliteit en query expansie

- SRU/CQL wordt gebruikt als zoek protocol en query taal met Clarin specifieke extenties
  - Hiermee wordt voortgebouwd op samenwerking binnen CLARIN-EU

SRU/CQL

- een standaard XML-georienteerd zoek protocol for Internet zoek vragen, gebruik makend van CQL (Contextual Query Language), a standaard syntax voor zoek vraag representatie.

- CQL specificeert 4 niveaus van ondersteuning (verkort)
  - **Level 0**: term based query + diagnostiek
  - **Level 1**: verdere syntax ondersteuning + diagnostiek
  - **Level 2**: volledige syntax ondersteuning + diagnostiek

Kandidaat centra

Provider				
MI	Mimore	Morphosyntactic variation databases	Dutch dialects	online, level 0, open issues,
MPI	IMDI-subset	Acquisition, ELD and Sign Language	Mulilingual	online, level 0, open issues,
INL	Lexica and Corpora	Lexica and Corpora	Dutch	online, level 0, open issues,
DANS	Theatre data	18th century newspaper and playlists	Dutch	Online, level 0, open issues
ICLTT	C4	Historical text corpus	German	Online, level 0, open issues
UPF	El Pais newspaper corpus	Text corpus	Catalan	Online, allows index-based queries
UniTub	Gutenberg	Text Corpus	German	planned
…	…	…	…	…

Deelnemende kandidaat centra internationaal

- Nationaal:
  - MI, MPI, DANS en INL

- Internationaal (CLARIN-EU):
  - ICLTT (Austria) , UPF (Barcelona) , UniTueb (Germany), OTA (United Kingdom), BAS (Germany), IDS (Germany), Gothenborg Uni (Sweden),

SRU/CQL

- SRU vereist ondersteuning van aantal verplichte operaties:
  - Search/Retrieve - zoeken
  - Scan – generieke verkorte browse lijst
  - Explain – beschrijving van service mogelijkheden

- CQL

  searchClause ::= index relation [/modifier] searchTerm

  index ::= [‘cmd.’] cmdIndex | [‘css.’] contentIndex

Query =
cmd.Organization isa Univerisity and dc.title any Liebe
and
css.word=Herz prox/s/1 css.lemma= zerreißen

SRU

Pid = "http://hdl.handle.net/11858/00-395C-0000-0000-6F63-1

<sru:recordPacking>XML</sru:recordPacking>
<sru:recordData>
<css:Resource xmlns:css="http://clarin.eu/ContentSearch" pid="http://hdl.handle.net/11858/00-395C-0000-0000-6F63-1">
<css:ResourceFragment pid="gtrp:300">
<css:DataView type="kwic">koe-e / ku.6 / kuːə</css:DataView>

<css:DataView type="kml">

css:DataView type=”kwic”

css:Metadata

```
<ccs:Metadata>
   <ccs:f key="page">unknown</ccs:f>
   <ccs:f key="date">1923</ccs:f>
   <ccs:f key="textType"/>
   <ccs:f key="pubRegion"/>
   <ccs:f key="author">Mayreder, Rosa</ccs:f>
   <ccs:f key="keywords”/>
   <ccs:f
key="availability">r_s_restricted_show</ccs:f>
   <ccs:f key="textClassLong">Sachtext</ccs:f>
</css:Metadata>
```

nates>

css:DataView type=”kml”

<translation>a cow</translation>
<location>A001p</location>
<placename>Midsland / Midslân</placename>
<latitude>53.3942864195</latitude>
<longitude>5.30243930434</longitude>
<tags>N(onbek,3,s)</tags>
<audioserver>http://dss01.meertens.knaw.nl/gtrp</audioserve
<soundfile></soundfile>
<starttime></starttime>
<endtime></endtime>
</metadata>
</data>
</css:DataView>

css:DataView type=”text/xml”

Simple UI for testing purposes

Welcome to Searching

Welcome to Searching

The CGN-Corpus (Corpus Gesprok
ESF corpus
IFA corpus
Childes corpus
Talkbank corpus
Gysseling
GTRP
DynaSand
DIDDD
C4 corpus
cobreix
com
comarcal
comarcals
comarques
començа
començar
competidor
complementaris
Compta

kind Search

[Click Here]
[Click Here]
[Click Here]
[Click Here]
[Click Here]
[Click Here]
[Click Here]
[Click Here]
[Click Here]
[Click Here]
[Click Here]

>

op niet [[kinderen]] ingesteld eigenlijk

andere geen [[kindertjes]] Maar ja

je voor [[kinderen]] Of andere

n kinderen hebben?

inderen. die zullen toch niet allemaal hier

hier wonen? dat zal toch niet waar zijn?

Semantische interoperabiliteit basis idee

Query:

Actor.Name any Peter

+ relaties (DCR + RR):

#sameAs (#Actor, #Person)
#sameAs (#Name, #FullName)

= geexpandeerde query:

Actor.Name any Peter
OR Actor.FullName any Peter
OR Person.Name any Peter
OR Person.FullName any Peter

From: SMC4LRT Semantic Mapping Component forLanguage Resources and Technology
2011-06-06 Matej Ďurčo, ICLTT, Vienna;

Semantische interoperabiliteit basis idee

Binnen (metadata) schema zijn ver
data categ

```
<xs:element name="Title" maxOccurs=
   <xs:complexType>
      <xs:simpleContent>
         <xs:extension base="xs:string"
            <xs:attribute ref="xml:lang"/>
         </xs:extension>
      </xs:simpleContent>
   </xs:complexType>
</xs:element>
```

Relation Regi
met andere

#sameAs

Data Category: project name

Key	2536
PID	http://www.isocat.org/datcat/DC-2536
Type	complex/open
Owner	Athens Core
Scope	public

1. Administration Information Section

1.1 Administration Record	
Identifier	projectName
Version	1:0
Registration Status	private
Administration Status	private
Justification	common metadata data category
Origin	IMDI: session.project.name
Effective Date	2009-08-13
1.1.1 Creation	
Creation Date	2009-08-13
Change Description	A short name or abbreviation of the project that led to the creation of the resource or tool/service.
Last Change	

Data Category: project title

cat.org/datcat/DC-2537

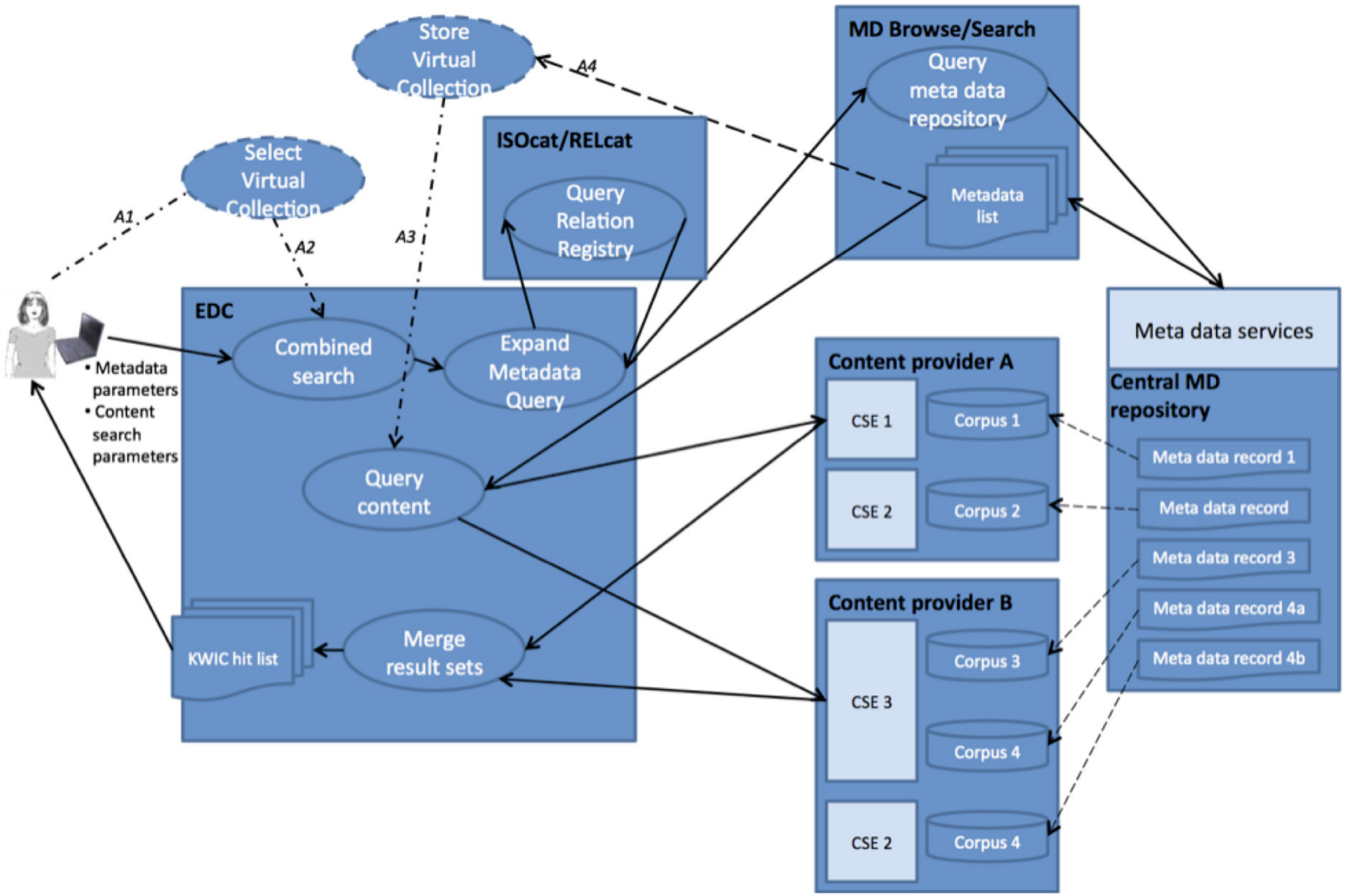
formation Section

d

data data category
.project.title

f the project that led to
the resource or

Technische opzet
Integratie van metadata search



Ontwikkelingen

- Specificatie van dataviews
- Stabilisatie van resultaat formaat
- Uitbreiding van zoek functionaliteit naar hoger compliance level (Diagnostics!)
- Manier om zoek engines te koppelen aan resources/metadata
- Integratie content search met metadata search
- User Interface !!!

- Uitbreiding van aantal zoek engines
- Gevanceerde topics zoals resultaat integratie en ranking

Thank you for your attention