



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Facets of a glimmering gem

Overview

Introduction

Litteraturbanken.se

Clarín JMDR

Indi use

Some (other)
uses

Final Comments

Leif-Jöran Olsson

Språkbanken, Department of Swedish Language, University of
Gothenburg

2010-03-12



Overview

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

- ▶ [Litteraturbanken.se](#)
- ▶ [Clarín.eu](#)
- ▶ [Some \(other\) uses](#)
- ▶ [Final comments](#)

Overview

Introduction

Litteraturbanken.se

Clarín JMDR

Final use

Some (other)
uses

Final Comments



Introduction

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

- ▶ Språkbanken
- ▶ Experience check – anyone here with a connection to Sweden or the Swedish language?
- ▶ Litteraturbanken.se
- ▶ Clarin Joint metadata repository (CJMDR)

Overview

Introduction

Litteraturbanken.se

Clarin JMDR

Imdi use

Some (other)
uses

Final Comments

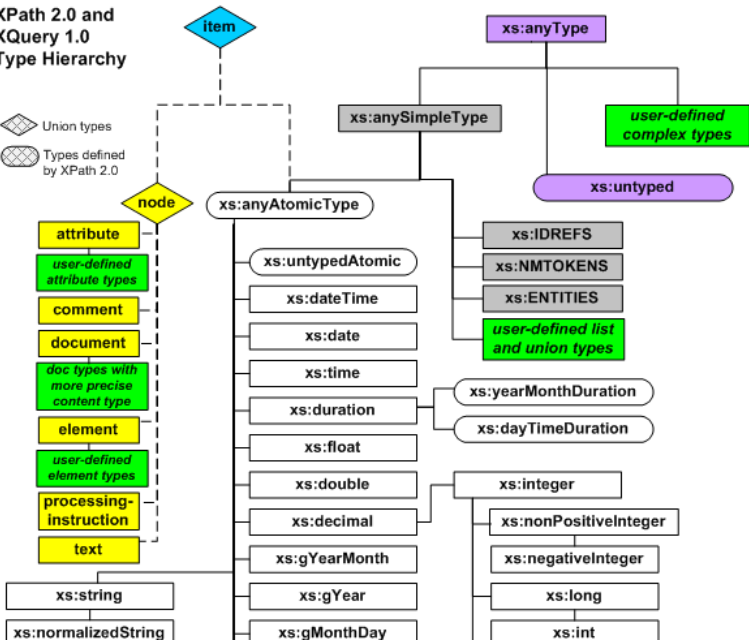


Data model

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

XPath 2.0 and XQuery 1.0 Type Hierarchy



Overview

Introduction

Litteraturbanken.se

Clarín JMDR

Imdi use

Some (other)

uses

Final Comments



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Litteraturbanken.se – purpose

Litteraturbankens syfte (purpose) är att vara en fri (free) kulturhistorisk och litterär resurs för forskning, undervisning (education) och folkbildning (life-long-learning). Den är till för alla: forskare, lärare, studerande och litterärt allmänintresserade. Huvuduppgiften är att samla in (collect) och digitalisera skönlitteratur (fiction) och viktigare humaniora samt tillgängliggöra (provide) materialet på sådant sätt att det blir möjligt för användare att arbeta med det.

Overview

introduction

Litteraturbanken.se

Clarín JMDR

Imdi use

Some (other)
uses

Final Comments



GÖTEBORGS
UNIVERSITET

Språk

BANKEN

Litteraturbanken.se – goal

Målet är att utveckla Litteraturbanken till en webbplats som rymmer det viktigaste inom svensk litteratur.

Overview

Introduction

Litteraturbanken.se

Clarin JMDR

Initial use

Some (other)
uses

Final Comments



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Litteraturbanken.se – some figures

Currently

- ▶ 413 works in total
- ▶ of which 249 are e-text works
- ▶ Rate of increase is around 120 works/year or above 12 per month since we only release 10 times per year.
- ▶ 45 GB of editor supplied material of which 38 GB are facsimilies (in 3-5 sizes).
- ▶ 16.3 million words in the e-texts
- ▶ 300.000+ named entities
- ▶ 5.600 lines of XQuery

Overview

introduction

Litteraturbanken.se

Clarín JMDR

Imdi use

Some (other)
uses

Final Comments

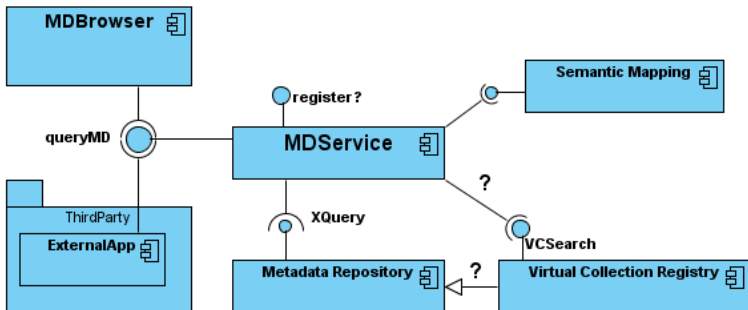


Clarín services overview

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

- Overview
- introduction
- Litteraturbanken.se
- Clarín JMDR
- Indi use
- Some (other) uses
- Final Comments





Interaction

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Overview

introduction

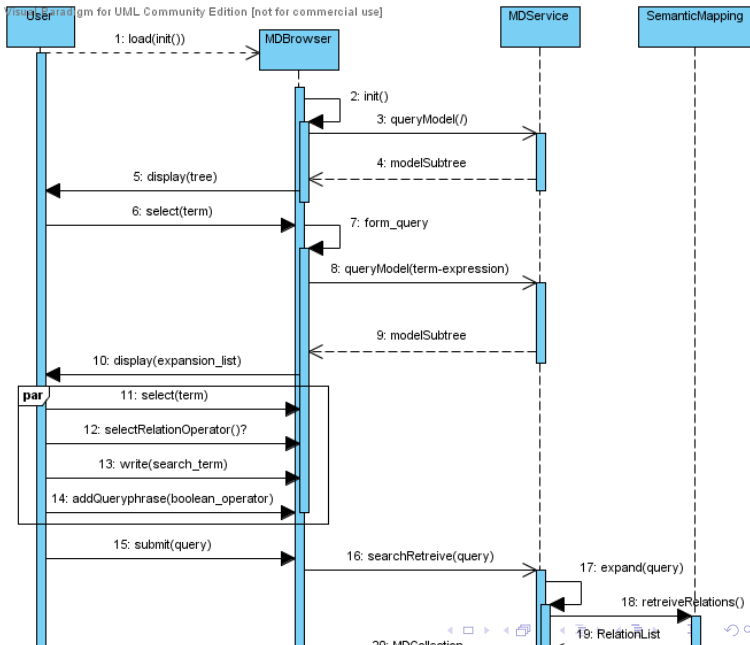
Litteraturbanken.se

Clarín JMDR

Imdi use

Some (other)
uses

Final Comments





Example

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Metadata - XPath

Metadata will primarily have to be converted to XPath for search in MDRepository

CLARIN QL	XPath	CLARIN QL Example	XPath Example
{cmdComponent}	//{cmdComponent}	Actor	//Actor
{cmdPath}	//{cmdPath}/{cmdComponent}	Actor.ContactPhone	//Actor/Contact/Phone
{cmdIndex} {rel} {term}	//{cmdIndex}[{rel} '{term}']	Actors.Actor.Sex=f	//Actors/Actor/Sex[=f]
{cmdIndex} any {term}	//{cmdIndex}[contains('{term}')]	Organisation Name any University	//Organisation/Name[contains(, 'University')]
and, or, and not	?!	Organisation Name any University and Actor.gender=m	?!

[Overview](#)

[Introduction](#)

[Litteraturbanken.se](#)

[Clarín JMDR](#)

[Imdi use](#)

[Some \(other\)
uses](#)

[Final Comments](#)



sign language corpus index (from IMDI)

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

[Overview](#)

[Introduction](#)

[Litteraturbanken.se](#)

[Clarín JMDR](#)

[Imdi use](#)

[Some \(other\)
uses](#)

[Final Comments](#)

The screenshot shows a web browser window displaying the content of an XML file named 'sign_language.imdi'. The XML content is as follows:

```
<?xml version="1.0" encoding="UTF-8"?>.  
<METATRANSCRIPT ArchiveHandle="hdl:1839/00-0000-0000-0004-DF8D-8" Date="2006-  
  <Corpus CorpusStructureService="" SearchService="">.  
  <Name>Sign Language</Name>.  
  <Title>Sign Language corpora</Title>.  
  <Description LanguageId="" Link=""/>.  
  <CorpusLink ArchiveHandle="hdl:1839/00-0000-0000-0001-494E-3" Name="E  
  <CorpusLink ArchiveHandle="hdl:1839/00-0000-0000-0004-DF8E-6" Name="C  
  <CorpusLink ArchiveHandle="hdl:1839/00-0000-0000-0001-2A9A-4" Name="S  
  <CorpusLink ArchiveHandle="hdl:1839/00-0000-0000-0004-DF8F-4" Name="V  
  <CorpusLink ArchiveHandle="hdl:1839/00-0000-0000-0008-68DB-A" Name="V  
  </Corpus>.  
</METATRANSCRIPT>.
```

Below the browser window, a terminal window shows the following commands and their output:

```
exist:/db/clarin-mpi/corpus1.mpi.nl/qfs1/media-archive/Corpusstruc  
ture/biling_data> cd "Corpusstructure"  
exist:/db/clarin-mpi/corpus1.mpi.nl/qfs1/media-archive/Corpusstruc  
ture/biling_data/Corpusstructure> cd ..  
exist:/db/clarin-mpi/corpus1.mpi.nl/qfs1/media-archive/Corpusstruc  
ture/biling_data> cd ..  
exist:/db/clarin-mpi/corpus1.mpi.nl/qfs1/media-archive/Corpusstruc  
ture>
```

The terminal window title is 'exist-administrationsklienten uppkopplad - admin@xmldb:exist://localhost...'. A 'Close' button is visible in the bottom right corner of the terminal window.



lexemes with mother fish (from SALDO)

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

[Overview](#)

[Introduction](#)

[Litteraturbanken.se](#)

[Clarín JMDR](#)

[Imlå use](#)

[Some \(other\)
uses](#)

[Final Comments](#)

Query Dialog

Query Input:

History: 1. </a a="1">[@a >= 0 and @a <= 2]
//lexem[mor[. eq "fisk..1"]].

Context: /db/saldo Display max.: 100

Results:

XML Trace

```
<lexem xml:id="abborre..1">.  
  <gf>abborre</gf>.  
  <mor>fisk..1</mor>.  
  <far>PRIM..1</far>.  
  <lemma>abborre..nn.1</lemma>.  
</lexem>.  
<lexem xml:id="benfisk..1">.  
  <gf>benfisk</gf>.  
  <mor>fisk..1</mor>
```



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Entry for fish (from DALINS ORDBOK)

http://litteraturbanken.se/query/dalin.xql?word=fisk&limit=10

Google

ookmarks startuppsidor Gmail Bugs: lb-new Fedora Planet Builds eXist - trunk eXist - ReleaseTodo

FISK

m. 2.

1) (nat. hist.) Rygggradsdjur med rödt, men kallt blod, utan lungor; andas förmedelst gälar och lever i vattnet, hvarti han fortskaffar sig (simmer) medelst så kallade simfenor. En stor, liten f. Rensa f. (Talesätt) Frisk, qvick, munter, liflig som en f. li vattnet, i hög grad. F`sina f-ar varma, fjällade, få påskrifvet, få tilltal, bannor, snäsor, snubbor. Hvarken fågel eller f., hvarken det ena eller andra. F-en vill gå i vatten, säges skämtvis, då man äter fisk och vill påminna om att tag "en sup på fisken", som det kallas. (Ordspr.) I lugnaste vatten, gå största f-arne, en lugn yta döjer ofta de våldsammaste lidelser. —

2) Brukas i sing. o. obestämd form, kollektivt, för att utmärka: a) Fiskar i allmänhet. Färsk, salt f. Fånga f.; -

b) Maträtt af fångad och kokt, stekt eller på annat sätt tillredd fisk. Hafva f. till förrätt. Äta f. —

3) (skepp.) Hvarje håll i däck för en mast, pump, o. s. v. —

4) (astr.) F-arne:

a) benämning på en stjernbild li djurkretsen. —

b) Ett af tecknen i djurkretsen. — Ss. *F-art*, *-fena*, *-fett*, *-fjäll*, *-gäl*, *-handel*, *-handlare*, *-handlerska* ell. *-månglerska*, *-hufvud*, *-smak*, *-sori*, *-stjert*, *-land*, *-tran*, *-yngel*.

Overview

Introduction

Litteraturbanken.se

Clarín JMDR

Imdi use

Some (other)

uses

Final Comments



GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Sentence count for corpus konkplus (from SBKHS)

Query Dialog

Query Input:

```
History: 1. let $a := <a a="1" />return $a[@a >= 0  
count(//item).
```

Context: /db/sbkhs-data/punkt Display max: 100

Results:

XML Trace

```
11412393.
```

[Overview](#)

[introduction](#)

[Litteraturbanken.se](#)

[Clarín JMDR](#)

[Imlá use](#)

[Some \(other\)
uses](#)

[Final Comments](#)



Final Comments

GÖTEBORGS
UNIVERSITET

Språk
BANKEN

Overview

Introduction

Litteraturbanken.se

Clarin JMDR

Imdi use

Some (other)
uses

Final Comments

▶ partitioning

▶ facets