Deliverable 3.1
Metadata Integration Report

DAM-LR

011841

Distributed Access Management
for
Language Resources

implemented as
Specific Support Action

Contract Number: *011841*

Project Coordinator: Peter Wittenburg

Project Web-Site: www.mpi.nl/dam-lr/

Deliverable: D3.1

Authors: Lund, MPI

Responsible: Lund

Date: 30.1.2007

Content

1 Introduction

One of the major points of the integration of the different archives is metadata interoperability. Within the DAM-LR project this has been achieved by adapting the existing metadata infrastructures to the [IMDI] metadata concepts and framework. The metadata framework plays a crucial role in the development of the distributed solution and therefore it's important that all repositories at least provide an interface to supply correct IMDI metadata descriptions. The IMDI descriptions are formalised by the IMDI XML Schema [XMLSCHEMA] version 3.0.
Since each partner can have their own local metadata infrastructure it's clear that every archive needs a semantic mapping from existing metadata descriptions and structures to the overall IMDI framework. These processes are described in the following chapters.

1.1 Corpus Structure

Most corpora are created in a tree-structure reflecting an initial categorisation of the language resources. These kinds of structures are supported by IMDI in the form of corpus nodes. One corpus node can contain a set of links to other corpus nodes or to a set of resource bundles. Each node has a name and other metadata elements to describe that point in the corpus structure.

1.2 Resource Bundles

The concept of resource bundles (or Sessions) is used as a container for all information about the circumstances and conditions of the linguistic event, groups the resources belonging to this linguistic event, records the administrative information of the event and describes the content of the event. An extensive description can be found in the IMDI documentation Metadata Elements for Session Descriptions [SESSIONS].

1.3 Data encodings

Although [XML] enables the use of different character encodings it's more practical for tools to use a single encoding within the IMDI framework. It was agreed that all data encodings will be [UNICODE]. This is supported by IMDI by using UTF-8 [RFC2279] to implement the encoding of UNICODE characters.

1.4 Archive growth during 2006

Lund Metadata
At Lund already some part of the archive is described by using early versions of IMDI metadata. The data and metadata of these four sub-corpora are analysed and made available for integration into the distributed IMDI domain.

1.5 Lund Corpus

The following archives are involved:

- Phonogrammarchiv Vienna, Austria (pha)
- Student English Corpus of Uppsala University (use)
- Swedia 2000, subpart of a research corpus of Swedish dialects (swedia)
- The Swedish part of the Spencer project (spencer)

A special case is the Helsinki corpus containing IMDI corpus files and a 'creative' mix of Corpus/Session files. These files should have been IMDI Catalogue files. Communication with the authors is needed to do custom modifications for a complete upgrade to IMDI 3.0.

The table below gives an overview of the amount of IMDI and resource files in the four sub-corpora. The first number is the actual files copied. The second number (between braces) is the number of linked files. A difference between these numbers indicates that not all the linked files could be copied. This happens when a link is broken or when it's not allowed to access the file (read permission).

IMDI overview

type	pha	use	swedia	spencer
imdi	112 (117)	1957 (1957)	313 (313)	50 (50)
imdi session	96	1489	224	40
imdi corpus	6	468	89	10
description files	2	1	1	4

Resource overview

format	pha	use	swedia	spencer
wav	0 (93)	-	223 (224)	-
gif	-	-	438 (448)	-
jpeg	-	-	116 (224)	-
txt	63 (66)	1489 (1489)	-	-
html	-	-	1	-
doc	-	-	-	0 (44)
cha	-	-	-	0 (40)
pdf	-	0 (1)	-	-
unknown	1 (4)	-	-	-

1.6 Metadata upgrade

To upgrade the Lund IMDI files to IMDI 3.0 several generic repairs and a few custom modifications were needed. A serious problem with the older IMDI files was that they contain characters encoded in UTF-8 mixed with characters in [ISO8859-1] and sometimes even Windows specific codes [WINLATIN1]. A tool was made to correct these non-UTF8 codes.
Generic modifications include upgrading IMDI 1.x to 3.0 and correcting all 3.0 files with available tools to make the IMDI files valid. Some custom corrections were done to synchronise the closed vocabulary values and correct syntax errors.

1.7 File types and formats

The resources in the Lund archive are correctly specified using IMDI Type/Format values. One exception is the Spencer corpus which uses the format 'text/doc' which is not a known MIME type and should be corrected.

1.8 Language identification upgrade

Language names and identifiers are based on the 14th edition of the Ethnologue list from SIL. An upgrade to the new list (15th edition, ISO 639-3) can be expected in the near future. The table below gives an indication of the changes to be made.

Edition 14		Edition 15	
Name	**Code**	**Name**	**Code**
Serbo-Croatian	SRC	Bosnian	bos
Serbo-Croatian	SRC	Croatian	hrv
Serbo-Croatian	SRC	Serbian	srp
Armenian	ARM	Armenian	hye
Cantonese	YUH	Chinese, Yue	yue
Kurdi	KDB	Kurdish, Central	ckb
Kurdi	KDB	Kurdish, Southern	sdh
Norwegian, bokmal	NRR	Norwegian, Bokmål	nob

2 INL Metadata

2.1 CGN Corpus

Recently, a 2nd version of the CGN corpus has been released. Information for this version has been added to the tables below.

IMDI overview

file type	CGN	CGN 2.0
imdi	12,893	12,907
imdi session	12,767	12,780
imdi corpus	126	126
description files	1	1

Resource overview

format	CGN	CGN 2.0
wav[1]	12,767	12,780
bpt[2]	14,199	14,212
lxk	12,767	12,780
pri	12,767	12,780
prx[3]	796	796
skp[4]	26,966	26,992
tag	12,767	12,780
tig[5]	1,303	1,303
eaf[6]	12,767	12,780

2.2 Metadata upgrade

The 1st version of the Dutch Spoken Corpus (or CGN) at the INL uses an older version of IMDI and is therefore the first choice to be integrated. The upgrade tool from IMDI 1.x to 3.0 is used to make the metadata valid according to the new schema and also corrections of cv values had to be done.

2.3 File types and formats

The format specifier of several text resources are upgraded to conform to the most recent IMDI format for written resources which is based on MIME types. The table below gives an overview of this conversion.

old value	new value
text/x-bpt	text/x-cgn-bpt+xml
text/x-lxk	text/x-cgn-lxk+xml
text/x-pri	text/x-cgn-pri+xml
text/x-prx	text/x-cgn-prx+xml
text/x-skp	text/x-cgn-skp+xml
text/x-tag	text/x-cgn-tag+xml
text/x-tig	text/x-cgn-tig+xml

[1] Although there are 12,780 wav files in the CGN, only 12,767 had annotations and metadata in the 1st version. CGN 2.0 contains all available annotations and metadata.
[2] There are 12,780 auto-generated bpts in CGN 2.0. 1,432 were manually verified and stored separately.
[3] The prx files contain manually verified prosodic annotations. Two (groups of) annotators annotated (the same) 398 resources.
[4] The skp files link annotations to the correct time frames in the sound files. Skps were created for three sets of annotations: 12,780 + 12,780 + 1,432.
[5] Syntactic annotations were created for 1,303 fragments.

2.4 Future plans

The CGN is the INL's test case. When the CGN has been integrated into the distributed DAM-LR solution, more language materials will follow.

3 SOAS Metadata

3.1 Update 2006

Due to some problems in staffing and archive software stability, a new setup was chosen begin 2007, so that a first collection from SOAS will be visible in the joint metadata domain. This work needs to be intensified in 2007. For more details we refer to the annual report.

3.2 Earlier State

The Endangered Languages Archive [ELAR], SOAS, has possibly the largest variety of potential contributors and users of the archive materials, and so the archive design and metadata policies are designed to provide the greatest possible flexibility in materials and associated metadata, while conforming to a large number of metadata standards.
After an extensive review of digital archive design principles, an archive architecture was designed that allows the maintainence a local metadata set that is continuously extensible, while serving and IMDI compliant version of the metadata in order to meet the requirements of the DAM-LR project.
The archive design is heavily influenced by the Open Archive Information System [OAIS] and in particular the division between ingestion (accession), archive and dissemination formats. The OAIS model defines sets of data in terms of 'information packages', which is defined as a bundle of resources plus the associated metadata. For ingestion, archiving and dissemination packages are respectively known as a Submission Information Package (SIP), Archival Information Package (AIP) and a Dissemination Information Package (DIP).

3.2.1 Ingestion metadata

The following metadata elements are the minimum required for accessioning materials into ELAR:

- Identifier: A means to uniquely identify each item in the SIP. This might be either:
  - a unique name for each item listed together with the full filename (and media carrier label if relevant) or
  - a unique filename for each item
- Format: Describe formats
  - file format
  - mark-up format
  - character encoding format
- Creator: Entity primarily responsible for making the content
- Subject.language: The language(s) which is described or documented
- Language: The language in which the content is expressed or introduced.
- Rights: Information about rights held in and over the resource (ELAR will apply default values if required)

Additionally, ELAR is developing dedicated applications for the automated ingestion of data in IMDI and [OLAC] compliant formats.

3.2.2 Archive metadata

ELAR encourages users to add metadata and new metadata categories. Therefore, a primary role in managing the metadata is moderating the ongoing development of metadata supplied by users. The AIP metadata is initiated with fields primarily taken from the IMDI metadata schema, with some initial extensions.
ELAR will be the first language archive to actively encourage users to translate metadata into different languages, to support the browsing of metadata via these different languages and to allow multiple concurrent values for metadata fields. This requires individual metadata values to be uniquely and unambiguously identifiable, which is a more fine-grained model of metadata than that which is supported by most current metadata schemas. While this doesn't prevent mapping archive metadata to well-known standards, is does prevent ELAR being 'IMDI-native', meaning that our metadata integration strategy for the DAM-LR project is to dynamically map our archive metadata to an IMDI compliant format.

3.2.3 Dissemination metadata

ELAR's catalogue serving system is part of the dissemination strategy, serving DIPs with metadata in IMDI, OLAC (Open Language Archives Community), OAI (Open Archives Initiative) and Text Encoding Initiative [TEI] formats. The AIP metadata will be dynamically mapped to each of these formats in order to serve a variety of user communities in parallel. An advantage of this mapping system is that the archive will be able to map to future versions of metadata formats without needing to change the archive metadata, thereby saving considerable future resources and the need to re-verify the changes for preservation purposes. Further details of the mappings to OLAC, OAI, and TEI formats are outside the scope of the DAM-LR project.

The mapping to IMDI is to meet the requirements of the DAM-LR project and to support users of the archive who wish to take advantage of IMDI compliant software tools. As the AIP metadata system has been initiated with IMDI fields, the mapping to IMDI metadata in the DIP metadata has been a simple task, and has not presented any significant design problems.

4 MPI Metadata

4.1 Archive Growth in 2006

The DAM-LR archive grew from 15 TB in 2005 to 25 TB 2006. Also the amount of available metadata descriptions increased correspondingly. All metadata is openly available via the archive web site.

4.2 General Work in 2006

At MPI much work was undertaken to increase the consistency of the metadata set and to add semi-automatically additional information. As an example for consistency improvement we can refer to the work on file format issues. Here we first developed and improved a number of parsers for different file types such as EAF/CHAR/Shoebox/Text annotations and media files which allow us to check whether the specified file extensions are correct. Almost all file types in the archive were corrected with the help of this checking process. Then all MIME type specifications in the metadata descriptions were changed appropriately. These time consuming operations have been finished now.

After having made extensive metadata statistics on the whole archive (see appendix B) we found out that the language names were filled in for almost all resources where it is was possible, however, the formal language code that can be used for searching, for example, was only used partly. Also in a time consuming semi-automatic process all names were compiled and compared. Based on the results decisions were taken and the language codes according to the Ethnologue standard (which will become an ISO standard) were added.

4.3 Earlier State

The last two years the major part of the MPI archive was upgraded from IMDI 1.x to 3.0. Also the diverse domain of resource formats was synchronized with standard formats and a clean-up of cv values was done. However, there are still a lot of resources to be processed and integrated which is an ongoing task of corpus management. Tools are continuously developed to assist corpus managers with checking and correction of new resources to be ingested into the archive.

4.3.1 MPI Corpus

The table below gives a global overview of the IMDI archive from last year.

IMDI overview

type	count
imdi	12893
imdi session	22004
imdi corpus	3576
description files	8024

Media Resources

format	count
jpg	8987
mov	3357
mpeg	1226
mpg	11543
wav	7909

Text Resources

format	count
chat	3323
eaf	1877
pdf	260
sht	387
tr	723

txt	5833
html	4
typ	8

4.3.2 DBD Corpus

A new IMDI session profile (see Appendix A) was made to enable IMDI metadata creation for the Dutch Bilingual Database (DBD). The corpus was added to the IMDI domain.

The following additional keys were defined:

Name	CV
DBD.LanguageMode	http://www.mpi.nl/IMDI/Schema/DBD.LanguageMode.xml
DBD.CountryOfBirth	http://www.mpi.nl/IMDI/Schema/Countries.xml
DBD.AgeAtImmigration	http://www.mpi.nl/IMDI/Schema/DBD.AgeAtImmigration.xml
DBD.LevelOfBilingualism	http://www.mpi.nl/IMDI/Schema/DBD.LevelOfBilingualism.xml

IMDI overview

type	count
imdi	1366
imdi session	1190
imdi corpus	176
description files	633

Media Resources

format	count
jpg	4
pdf	5
wav	194

Text Resources

format	count
chat	668
eaf	3
pdf	21
txt	1

5 References

[ELAR] Endangered Languages Archive
http://www.hrelp.org/archive/

[IMDI] ISLE Metadata Initiative
http://www.mpi.nl/IMDI/

[ISO8859-1] ISO-8859. International Standard, Information Processing, 8-bit Single-Byte Coded Graphic Character Sets, Part 1: Latin alphabet No. 1, 1987

[OAI] Open Archives Initiative
http://www.openarchives.org/

[OAIPMH] The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0, 2002-06-14
http://www.openarchives.org/OAI/openarchivesprotocol.html

[OAIS] Reference model for an Open Archival Information System, January 2002
http://www.ccsds.org/documents/650x0b1.pdf

[OLAC] Open Language Archive Community
http://www.language-archives.org/

[RFC1738] Uniform Resource Locators (URL), December 1994
http://www.ietf.org/rfc/rfc1738.txt

[RFC2279] UTF-8, a transformation format of ISO 10646, January 1998
http://www.ietf.org/rfc/rfc2279.txt

[SESSIONS] Metadata Elements for Session Descriptions, Version 3.0.4, MPI Nijmegen, 2003
http://www.mpi.nl/IMDI/

[TEI] Text Encoding Initiative
http://www.tei-c.org/

[UNICODE] Unicode Standard
http://www.unicode.org/

[WINLATIN1] Windows-1252 codepage (WinLatin1)
http://en.wikipedia.org/wiki/Windows-1252

[XML] Extensible Markup Language (XML) 1.0 (Third Edition), W3C, 2004
http://www.w3.org/TR/REC-xml/

[XMLSCHEMA] XML Schema Part 1: Structures Second Edition, W3C, October 2004
http://www.w3.org/TR/xmlschema-1/

Appendix A : IMDI DBD Session Profile

```
<?xml version="1.0" encoding="UTF-8"?><!DOCTYPE METATRANSCRIPT [<!ENTITY
annotationunitPrefix "">
<!ENTITY infolinkPrefix "">
<!ENTITY globalPrefix "">
<!ENTITY anonymousPrefix "">
<!ENTITY mediafilePrefix "">
]>
<METATRANSCRIPT Date="2005-06-17" FormatId="IMDI 3.0" Originator="Editor -
Profile:local/DBD_Profile.Profile.xml" Type="SESSION.Profile" Version="1"
xmlns="http://www.mpi.nl/IMDI/Schema/IMDI"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI  ./IMDI_3.0.xsd">
    <Session XXX-HelpText="A session bundles information about the cicumstances and
conditions of the linguistic event, groups the resources belonging to this
linguistic event, records the administrative information of the event, and
describes the content of the event." XXX-Multiple="false">
        <Name XXX-HelpText="A short name to identify the session. The name of the
session can be considered as shorthand of the session title." XXX-Multiple="false"
XXX-Type="String">DBD Profile</Name>
        <Title XXX-Type="String"/>
        <Date XXX-Type="Date">Unspecified</Date>
        <Description LanguageId="" Link="" XXX-Multiple="true" XXX-Type="String"/>
        <MDGroup XXX-Visible="false">
            <Location>
                <Continent Link="http://www.mpi.nl/IMDI/Schema/Continents.xml"
Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Continent>
                <Country Link="http://www.mpi.nl/IMDI/Schema/Countries.xml"
Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Country>
                <Region XXX-Multiple="true" XXX-Type="String">Unspecified</Region>
                <Address XXX-Type="String">Unspecified</Address>
            </Location>
            <Project>
                <Name XXX-Type="String"/>
                <Title XXX-Type="String"/>
                <Id XXX-Type="String"/>
                <Contact>
                    <Name XXX-Type="String"/>
                    <Address XXX-Type="String"/>
                    <Email XXX-Type="String"/>
                    <Organisation XXX-Type="String"/>
                </Contact>
                <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
            </Project>
            <Keys>
                <Key Name="" Type="OpenVocabulary" XXX-Multiple="true"/>
            </Keys>
            <Content XXX-Multiple="false">
                <Genre Link="http://www.mpi.nl/IMDI/Schema/Content-Genre.xml"
Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Genre>
                <SubGenre DefaultLink=""
Link="http://www.mpi.nl/IMDI/Schema/Content-SubGenre.xml" Type="OpenVocabularyList"
XXX-FollowUpDepend="Genre" XXX-Type="FixedCV">Unspecified</SubGenre>
                <Task Link="http://www.mpi.nl/IMDI/Schema/Content-Task.xml"
Type="OpenVocabulary" XXX-Type="FixedCV"/>
                <Modalities Link="http://www.mpi.nl/IMDI/Schema/Content-
Modalities.xml" Type="OpenVocabularyList" XXX-Type="FixedCV"/>
                <Subject DefaultLink=""
Link="http://www.mpi.nl/IMDI/Schema/Content-Subject.xml" Type="OpenVocabularyList"
XXX-Type="FreeCV"/>
                <CommunicationContext>
                    <Interactivity Link="http://www.mpi.nl/IMDI/Schema/Content-
Interactivity.xml" Type="ClosedVocabulary" XXX-
Type="FixedCV">Unspecified</Interactivity>
```

```
                    <PlanningType Link="http://www.mpi.nl/IMDI/Schema/Content-
PlanningType.xml" Type="ClosedVocabulary" XXX-
Type="FixedCV">Unspecified</PlanningType>
                    <Involvement Link="http://www.mpi.nl/IMDI/Schema/Content-
Involvement.xml" Type="ClosedVocabulary" XXX-
Type="FixedCV">Unspecified</Involvement>
                    <SocialContext Link="http://www.mpi.nl/IMDI/Schema/Content-
SocialContext.xml" Type="ClosedVocabulary" XXX-
Type="FixedCV">Unspecified</SocialContext>
                    <EventStructure Link="http://www.mpi.nl/IMDI/Schema/Content-
EventStructure.xml" Type="ClosedVocabulary" XXX-
Type="FixedCV">Unspecified</EventStructure>
                    <Channel Link="http://www.mpi.nl/IMDI/Schema/Content-
Channel.xml" Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Channel>
                </CommunicationContext>
                <Languages>
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                    <Language ResourceRef="" XXX-Multiple="true">
                        <Id XXX-Type="String">Unspecified</Id>
                        <Name Link="http://www.mpi.nl/IMDI/Schema/MPI-
Languages.xml" Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Name>
                        <Dominant XXX-Type="Boolean">Unspecified</Dominant>
                        <SourceLanguage XXX-
Type="Boolean">Unspecified</SourceLanguage>
                        <TargetLanguage XXX-
Type="Boolean">Unspecified</TargetLanguage>
                        <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                    </Language>
                </Languages>
                <Keys XXX-Multiple="false">
                    <Key DefaultLink="" Name="DBD.LanguageMode" XXX-
Multiple="false" XXX-
Type="CV:http://www.mpi.nl/IMDI/Schema/DBD.LanguageMode.xml">Unspecified</Key>
                </Keys>
                <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
            </Content>
            <Actors XXX-Multiple="false" XXX-Visible="false">
                <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                <Actor ResourceRef="" XXX-Multiple="true" XXX-Tag="Consultant">
                    <Role Link="http://www.mpi.nl/IMDI/Schema/Actor-Role.xml"
Type="OpenVocabularyList" XXX-Multiple="false" XXX-Type="FixedCV">Consultant</Role>
                    <Name XXX-Type="String"/>
                    <FullName XXX-Type="String"/>
                    <Code XXX-Type="String"/>
                    <FamilySocialRole Link="http://www.mpi.nl/IMDI/Schema/Actor-
FamilySocialRole.xml" Type="OpenVocabularyList" XXX-
Type="FixedCV">Unspecified</FamilySocialRole>
                    <Languages>
                        <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                        <Language XXX-Multiple="true" XXX-Type="Language">
                            <Id XXX-Type="String">Unspecified</Id>
                            <Name Link="http://www.mpi.nl/IMDI/Schema/MPI-
Languages.xml" Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Name>
                            <MotherTongue XXX-
Type="Boolean">Unspecified</MotherTongue>
                            <PrimaryLanguage XXX-
Type="Boolean">Unspecified</PrimaryLanguage>
                            <Description LanguageId="" Link="" XXX-Multiple="true"
XXX-Type="String"/>
                        </Language>
                    </Languages>
                    <EthnicGroup Type="OpenVocabulary" XXX-Type="String"/>
                    <Age XXX-Type="Age">Unspecified</Age>
```

```
                    <BirthDate XXX-Type="Date">Unspecified</BirthDate>
                    <Sex Link="http://www.mpi.nl/IMDI/Schema/Actor-Sex.xml"
Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Sex>
                    <Education XXX-Type="String">Unspecified</Education>
                    <Anonymized XXX-Type="Boolean">Unspecified</Anonymized>
                    <Contact>
                        <Name XXX-Type="String"/>
                        <Address XXX-Type="String"/>
                        <Email XXX-Type="String"/>
                        <Organisation XXX-Type="String"/>
                    </Contact>
                    <Keys XXX-Multiple="false">
                        <Key DefaultLink="" Name="DBD.CountryOfBirth" XXX-
Multiple="false" XXX-
Type="CV:http://www.mpi.nl/IMDI/Schema/Countries.xml">Unspecified</Key>
                        <Key DefaultLink="" Name="DBD.AgeAtImmigration" XXX-
Multiple="false" XXX-
Type="CV:http://www.mpi.nl/IMDI/Schema/DBD.AgeAtImmigration.xml">Unspecified</Key>
                        <Key DefaultLink="" Name="DBD.LevelOfBilingualism" XXX-
Multiple="false" XXX-
Type="CV:http://www.mpi.nl/IMDI/Schema/DBD.LevelOfBilingualism.xml">Unspecified</Ke
y>
                    </Keys>
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                </Actor>
                <Actor ResourceRef="" XXX-Multiple="true" XXX-Tag="other">
                    <Role Link="http://www.mpi.nl/IMDI/Schema/Actor-Role.xml"
Type="OpenVocabularyList" XXX-Type="FixedCV">Unspecified</Role>
                    <Name XXX-Type="String"/>
                    <FullName XXX-Type="String"/>
                    <Code XXX-Type="String"/>
                    <FamilySocialRole Link="http://www.mpi.nl/IMDI/Schema/Actor-
FamilySocialRole.xml" Type="OpenVocabularyList" XXX-
Type="FixedCV">Unspecified</FamilySocialRole>
                    <Languages>
                        <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                        <Language XXX-Multiple="true" XXX-Type="Language">
                            <Id XXX-Type="String">Unspecified</Id>
                            <Name Link="http://www.mpi.nl/IMDI/Schema/MPI-
Languages.xml" Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Name>
                            <MotherTongue XXX-
Type="Boolean">Unspecified</MotherTongue>
                            <PrimaryLanguage XXX-
Type="Boolean">Unspecified</PrimaryLanguage>
                            <Description LanguageId="" Link="" XXX-Multiple="true"
XXX-Type="String"/>
                        </Language>
                    </Languages>
                    <EthnicGroup Type="OpenVocabulary" XXX-Type="String"/>
                    <Age XXX-Type="Age">Unspecified</Age>
                    <BirthDate XXX-Type="Date">Unspecified</BirthDate>
                    <Sex Link="http://www.mpi.nl/IMDI/Schema/Actor-Sex.xml"
Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Sex>
                    <Education XXX-Type="String">Unspecified</Education>
                    <Anonymized XXX-Type="Boolean">Unspecified</Anonymized>
                    <Contact>
                        <Name XXX-Type="String"/>
                        <Address XXX-Type="String"/>
                        <Email XXX-Type="String"/>
                        <Organisation XXX-Type="String"/>
                    </Contact>
                    <Keys>
                        <Key Name="" Type="OpenVocabulary" XXX-Multiple="true" XXX-
Type="*"/>
                    </Keys>
```

```
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                </Actor>
            </Actors>
        </MDGroup>
        <Resources>
            <MediaFile XXX-Multiple="true">
                <ResourceLink XXX-Type="URL"/>
                <Type Link="http://www.mpi.nl/IMDI/Schema/MediaFile-Type.xml"
Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Type>
                <Format Link="http://www.mpi.nl/IMDI/Schema/MediaFile-Format.xml"
Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Format>
                <Size XXX-Type="String"/>
                <Quality XXX-
Type="Regexp:[12345]|Unknown|Unspecified">Unspecified</Quality>
                <RecordingConditions XXX-Type="String"/>
                <TimePosition>
                    <Start XXX-Type="Regexp:Unknown|[0-9][0-9]:[0-9][0-9]:[0-9][0-
9](:[0-9]+)?|Unspecified">Unspecified</Start>
                    <End XXX-Type="Regexp:Unknown|[0-9][0-9]:[0-9][0-9]:[0-9][0-
9](:[0-9]+)?|Unspecified">Unspecified</End>
                </TimePosition>
                <Access>
                    <Availability Type="OpenVocabulary" XXX-Type="String"/>
                    <Date XXX-Type="Date">Unspecified</Date>
                    <Owner XXX-Type="String"/>
                    <Publisher XXX-Type="String"/>
                    <Contact>
                        <Name XXX-Type="String"/>
                        <Address XXX-Type="String"/>
                        <Email XXX-Type="String"/>
                        <Organisation XXX-Type="String"/>
                    </Contact>
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                </Access>
                <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                <Keys>
                    <Key Name="" Type="OpenVocabulary" XXX-Multiple="true" XXX-
Type="*"/>
                </Keys>
            </MediaFile>
            <WrittenResource XXX-Multiple="true">
                <ResourceLink XXX-Type="URL"/>
                <MediaResourceLink XXX-Type="URL"/>
                <Date XXX-Type="Date">Unspecified</Date>
                <Type Link="http://www.mpi.nl/IMDI/Schema/WrittenResource-Type.xml"
Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Type>
                <SubType Link="http://www.mpi.nl/IMDI/Schema/WrittenResource-
SubType.xml" Type="OpenVocabularyList" XXX-FollowUpDepend="Type" XXX-
Type="FixedCV">Unspecified</SubType>
                <Format Link="http://www.mpi.nl/IMDI/Schema/WrittenResource-
Format.xml" Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Format>
                <Size Type="OpenVocabulary" XXX-Type="String"/>
                <Validation>
                    <Type Link="http://www.mpi.nl/IMDI/Schema/Validation-Type.xml"
Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Type>
                    <Methodology Link="http://www.mpi.nl/IMDI/Schema/Validation-
Methodology.xml" Type="ClosedVocabulary" XXX-
Type="FixedCV">Unspecified</Methodology>
                    <Level XXX-Type="Percentage">Unspecified</Level>
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                </Validation>
                <Derivation Link="http://www.mpi.nl/IMDI/Schema/WrittenResource-
Derivation.xml" Type="ClosedVocabulary" XXX-Type="FixedCV">Unspecified</Derivation>
                <CharacterEncoding XXX-Type="String"/>
```

```
                <ContentEncoding XXX-Type="String"/>
                <LanguageId Type="OpenVocabulary" XXX-Type="String"/>
                <Anonymized XXX-Type="Boolean">Unspecified</Anonymized>
                <Access>
                    <Availability Type="OpenVocabulary" XXX-Type="String"/>
                    <Date XXX-Type="Date">Unspecified</Date>
                    <Owner XXX-Type="String"/>
                    <Publisher XXX-Type="String"/>
                    <Contact>
                        <Name XXX-Type="String"/>
                        <Address XXX-Type="String"/>
                        <Email XXX-Type="String"/>
                        <Organisation XXX-Type="String"/>
                    </Contact>
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                </Access>
                <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                <Keys>
                    <Key Name="" Type="OpenVocabulary" XXX-Multiple="true" XXX-
Type="*"/>
                </Keys>
            </WrittenResource>
            <Source XXX-Multiple="true">
                <Id XXX-Type="String"/>
                <Format Link="http://www.mpi.nl/IMDI/Schema/Source-Format.xml"
Type="OpenVocabulary" XXX-Type="FixedCV">Unspecified</Format>
                <Quality XXX-
Type="Regexp:[12345]|Unknown|Unspecified">Unspecified</Quality>
                <CounterPosition>
                    <Start XXX-Type="Regexp:Unknown|[0-
9]+|Unspecified">Unspecified</Start>
                    <End XXX-Type="Regexp:Unknown|[0-
9]+|Unspecified">Unspecified</End>
                </CounterPosition>
                <TimePosition>
                    <Start XXX-Type="Regexp:Unknown|[0-9][0-9]:[0-9][0-9]:[0-9][0-
9](:[0-9]+)?|Unspecified">Unspecified</Start>
                    <End XXX-Type="Regexp:Unknown|[0-9][0-9]:[0-9][0-9]:[0-9][0-
9](:[0-9]+)?|Unspecified">Unspecified</End>
                </TimePosition>
                <Access>
                    <Availability Type="OpenVocabulary" XXX-Type="String"/>
                    <Date XXX-Type="Date">Unspecified</Date>
                    <Owner XXX-Type="String"/>
                    <Publisher XXX-Type="String"/>
                    <Contact>
                        <Name XXX-Type="String"/>
                        <Address XXX-Type="String"/>
                        <Email XXX-Type="String"/>
                        <Organisation XXX-Type="String"/>
                    </Contact>
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                </Access>
                <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                <Keys>
                    <Key Name="" Type="OpenVocabulary" XXX-Multiple="true" XXX-
Type="*"/>
                </Keys>
            </Source>
            <Anonyms>
                <ResourceLink XXX-Type="URL"/>
                <Access>
                    <Availability Type="OpenVocabulary" XXX-Type="String"/>
                    <Date XXX-Type="Date">Unspecified</Date>
```

```
                    <Owner XXX-Type="String"/>
                    <Publisher XXX-Type="String"/>
                    <Contact>
                        <Name XXX-Type="String"/>
                        <Address XXX-Type="String"/>
                        <Email XXX-Type="String"/>
                        <Organisation XXX-Type="String"/>
                    </Contact>
                    <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
                </Access>
            </Anonyms>
        </Resources>
        <References>
            <Description LanguageId="" Link="" XXX-Multiple="true" XXX-
Type="String"/>
        </References>
    </Session>
</METATRANSCRIPT>
```

Appendix B : IMDI Overview Paper

Comparison of Resource Discovery Methods

Alex Klassmann, Freddy Offenga, Daan Broeder, Romuald Skiba, Peter Wittenburg

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen
{alex.klassmann,freddy.offenga,daan.broeder,romuald.skiba,peter.wittenburg}@mpi.nl

Abstract

It is an ongoing debate whether categorical systems created by some experts are an appropriate way to help users finding useful resources in the internet. However for the much more restricted domain of language documentation such a category system might still prove reasonable if not indispensable. This article gives an overview over the particular IMDI category set and presents a rough evaluation of it's practical use at the Max-Planck-Institute Nijmegen.

1 Introduction

The raw material for linguists are samples of a particular language. These may range from pieces of parchment til recordings of TV broadcast. Although there exist guidelines for the metadata description and annotation of linguistic resources (IMDI [1], DC/OLAC [2], TEI [3], EAGLES [4], specialized data bases), no standard is universally accepted and probably can't be since researchers will focus on different aspects and invent new theories and ideas. The amount of collected and electronically available resources has exploded over recent years and poses the problem of organization/management and (re-)discovery of the data. In this paper we will present the approach the MPI for Psycholinguistics has chosen with respect to the metadata description, will elaborate on a number of different location methods and finally will discuss some critical points. The first paragraph will give a short overview over the IMDI metadata scheme. Than their practical application i.e. the tools which allow the user to handle this metadata set will be presented. A rough evaluation of the quality of the at present available metadata follows. Than an alternative to formal categorization will be presented, namely free „tagging“, which is currently lively discussed with respect to internet search engines. It's applicability to the field of linguistics will be questioned and some preliminary conclusions drawn.

2 IMDI Metadata

The IMDI (ISLE MetaData Initiative) scheme was developed during 2001-2003 by a broad network of linguists from different sub-disciplines such as field linguistics, phonetics, multimodality research and corpus linguistics. Its purpose is to give a solid, precise and extensible framework for the organization, bundling and retrieval of in principle any kind of digital linguistic resources, in particular annotated media streams and text sequences making up by far the largest percentage of current resources in language resource archives.

Typically primary language documents like audio or video files are accompanied by one or more text files, containing a transcription, translations and annotations at other linguistic levels (morphosyntax, semantic, etc) of the former and seen in the IMDI framework as resources themselves. An IMDI-session contains a detailed meta description of those tightly connected resources, and could therefore be named equivalently as metadata about a 'resource bundle'. The IMDI-schema describes in addition how those sessions can be grouped together into corpora and sub-corpora. Although corpus organization is relevant for management and browsing, it is not of relevance in this paper, i.e., for more details we refer to other IMDI documentation [5,6].

An IMDI-session can be best thought of as a form with roughly 150 hierarchically ordered entries, which concern e.g. information about

- the event (recording location, date, etc),
- the languages involved,
- the speaker(s),
- the type and nature of speech,

- technical information about the resources and
- access rights.

For most fields one or more values can be selected, but there are also so-called descriptive fields for the input of free text. Furthermore there is the possibility for every user to add arbitrary key-value-pairs which can be interpreted as a personal or project-specific extension of the schema. In order to facilitate the procedure of filling in the metadata, a special professional editor has been build at the Institute.

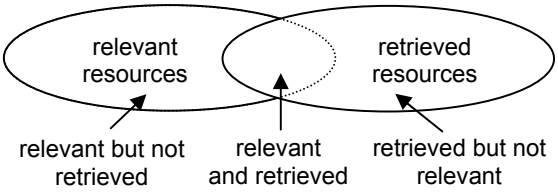
A single field, the „bundle name"-field is obligatory, yet users are urged to fill in all others, too. Unfortunately they tend to avoid this time-consuming work oriented to a re-usage by others and fields stay empty or have a default setting. Although everyone agrees that filling in metadata is very important in many respects, in particular since the knowledge about the content may be lost within shortest time, the amount of time spent on this aspect in the whole resource management life cycle is still too little.

3 Methodological Issues

One important question for the usage of archives – traditional as well as modern – with an extremely growing amount of resources is the possibility for the user to locate useful resources. As described the MPI uses the structured IMDI set to describe resources which therefore lends itself to carry out queries. Metadata includes added value with respect to the resources themselves, therefore it is data that cannot be missed. A recording may include an interview with a person having certain characteristics such as age, sex, education etc. Only in rare situations the recording will contain this information explicitly – it is the metadata description that will allow the interested user to make a comparison between male and female language use for example. Many other examples of this added value can be given.

Although we will have very different user groups ranging from researchers, teachers, students, journalists to the speakers themselves. All have different types of queries and all asking different types of interfaces. Nevertheless, we can make a few general statements on what a typical search method should optimize.

Literature defines two terms, "precision" and "recall", as measures for the success of a query. With "precision" the proportion of hits that are relevant compared to the irrelevant hits is meant. A higher amount of "noisy" results would therefore reduce the precision rate. With "recall" the proportion of relevant hits that were found compared to the not found relevant hits is meant. A query method that would not find very much of the relevant resources a user is looking for obviously would be not successful. The following drawing taken from G. Simons [7] is very useful to indicate the relation between the two terms.



Another important point in searching is of course the question of how to rank the hits. The precision could be very low, i.e., the number of irrelevant hits could be high, but if the relevant resources would be presented at the top of the list the user probably wouldn't bother. In this paper we will not discuss the ranking aspect.

4 The MPI Archive

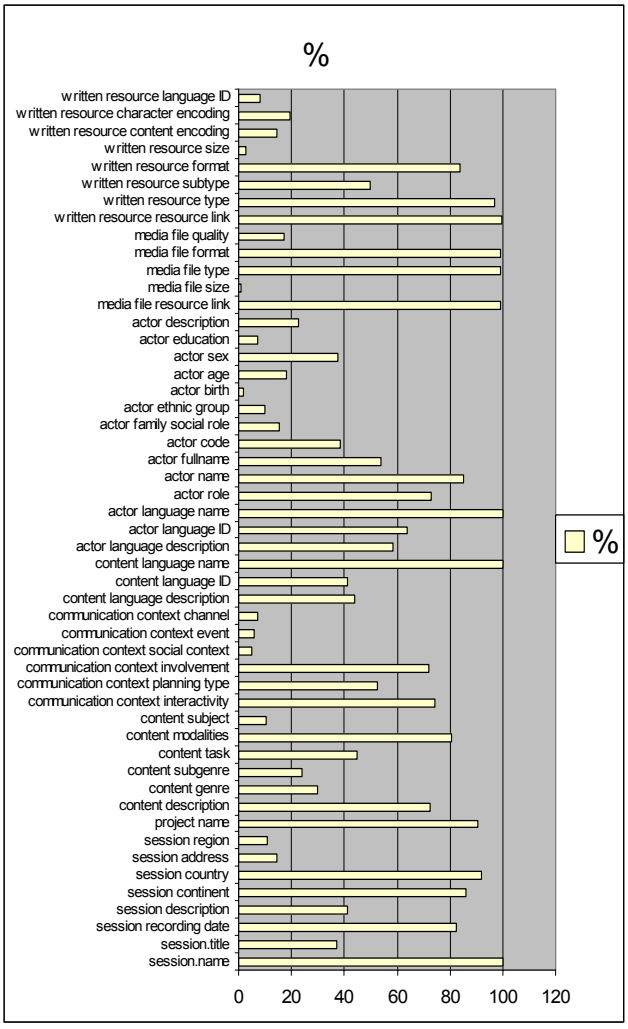
The Max-Planck-Institute Nijmegen houses a digital archive with a large variety of different language corpora, all categorized with the IMDI metadata set. The archive encompasses ca. individual 45.000 IMDI-sessions describing about 150.000 resources.

Infrastructure and tools have been designed to offer to the user several options to search for a specific IMDI-described resource. Since metadata is open per definition, all descriptions are accessible via the web; cf. http://corpus1.mpi.nl/ds/imdi_browser):

1) Browsing in linked resources. This is similar to clicking through a local file system with the difference that the hierarchy of corpus structures is much more stable. The approach is aimed at users familiar with or quickly able to grasp the underlying logical organization. Bookmarks help to make this process more efficient.

2) Structured search within the whole archive as well as within a selected part of it. Every IMDI-element can be addressed individually and the search for different elements can be combined into one query. Queries like "Give me all video files that show a female Wichita speaker older then 60 years" can be formulated and a high precision, i.e., a low number of irrelevant hits, can be expected. Yet, the user has to know the terminology used

by the IMDI schema in order to achieve a high recall, i.e., get a high percentage of the resources having looked for as hits. Furthermore, search is restricted to elements with closed or open vocabularies and does not cover elements with free text.



3) Unstructured search over the whole or part of the archive. The user can enter words or regular expressions into a free text field (Google-like). Any metadata element including the free text descriptions that contains matching strings will produce a hit. It is possible to formulate logical combinations of expressions and even "fuzzy terms" (for an overview of the possibilities cf. [8]). The recall with this method can be expected to be higher compared with structured search, however, the precision will be poor, i.e., much more irrelevant hits can be expected.

4) An extension of unstructured search is to provide the metadata descriptions to web search engines like Google with their advanced information retrieval techniques. However search cannot be restrained to a specific corpus, not to mention parts of it, and results will include a huge amount of unwanted hits from the whole internet. An additional term such as „IMDI“ or „MPI“ improves the precision significantly, but still yields unsatisfactory results.[7]

5) All IMDI records were transmitted to the OLAC service provider (DC [9]). OLAC offers a structured search possibility, but limits itself to the elements of DC and a few additional ones such as the language a resource is in. Currently, the service is not working well, since the OLAC service provider does not accept too many records, i.e., they expect the data provider to just deliver one metadata record for a sub-corpus. For the MPI it is in many cases difficult to determine what exactly a sub-corpus is. With respect to precision and recall we expect similar results as with structured search, as long as the restricted set of elements is sufficient. An advantage of using OLAC, however, is that other archives will contribute to OLAC, too.

6) Geographically orientated browsing. Since many languages in the archive are related to diverse and less known regions all over the world, a geographical browsing makes sense, too. The visualization tool Google-Earth [10] is used for this purpose, where the user can look for spots on the physical map of the Earth that point to IMDI-files. Of course, this method yields an enormous high precision and recall if only the geographic location is the discovery criterion. Since this approach is of less theoretical interest, we will not elaborate on this option.

We should not forget to mention that in general researchers want to combine metadata search/browsing with searching on the content as it is possible now for example with ANNEX [11]. Typical questions such as “give me all instances where a 4 year old female speaker is using a certain morpho-syntactic construction” can only be addressed when a combined structured search is performed. But we also understand that such questions will only be addressed by the “very well informed” user who knows exactly the terminology that is used. All other search options will not lead to useful results. In this paper we will not include the content search option, but discuss metadata search options in general.

5 Evaluation

In order to have significant variance in the data, an evaluation of the metadata was done on a subset of the resources in the archive, where metadata was filled in manually and by different users, i.e., the Dutch Spoken Corpus, for example, was not included.

[7] When searching for example for real resources for the TEOP language a Google search with “teop” as query string yields 17.600 hits with lots of unusable hits. A query string “imdi teop” only yields 683 hits and more important the entry for the Teop corpus is amongst the first five. However, users suffer from the same deficit: how should they know which string to use to achieve an acceptable precision and recall.

The table below gives an impression of how often fields are actually filled in (e.g. not empty and not default values like „unknown“ or „unspecified“). These statistics were created on 23.710 resource bundles. As can be seen the sets are far from being complete. On the other hand, every field of the scheme (including those not shown in the table) has been used in some sessions, so that it seems that no field in the schema is obsolete. These statistics give a baseline idea of what can be expected.

Since there is still not sufficient experience at the institute with actually performed metadata searches, it is not yet possible to carry out a full-fledged statistical evaluation based on empirical data. Instead, test queries which might be of relevance for researchers were formulated and executed. It was then checked whether the hits were accurate.

So, e.g. in Second Language Learning Research the influence of age on the acquisition of language is examined and it is assumed that there is a critical period in childhood for the development of certain skills such as learning grammar constructions. In order to find resources one would like to formulate a query like „Give me all resources for a given (not-mother-)language for speakers aged between 4 and 16 years“. Since the development between boys and girls may differ one even could refine the query by an appropriate qualifier.

Using the IMDI structured search the following query “Language=Dutch, Actor.Language.Mothertongue=false, Actor-Age<16 and >4“ yields 203 hits. An additional selection on “Actor-sex = Male“ results in 119 hits and one with “Actor-sex = Female“ in 83 hits. A full-text search with a query “Dutch AND second AND language AND (15 OR ... OR 5)“ results in 488 hits and may be still useful, too.

Categorization with respect to age and sex as well as technical categories like the file format are rather uncontested and not prone to subjective interpretation. This is different with respect to the descriptive elements concerning the content. Here the difficulty can be seen at the many corrections the initial IMDI set experienced and the user is merely offered a list of given values, but can type in others (“open vocabulary“).

The vocabulary for the element „Content-Genre“ e.g. encompasses 13 items („discourse“, „poetry“ etc.), two of them never have been used („Popular fiction“, „Newspaper article“) and another 15 values have been added by users. Concerning the element „Content-SubGenre“ the situation is similar: no offered type of drama has been used and (fortunately!) no resource was classified as „Unintelligible Speech“. Some 30 items were added, ranging from broad terms like „Speech“ to very specific ones. This poses the question if such a categorization in advance by a group of „experts“ is the right approach for data organization.

6 Free Tagging

In this paragraph we will discuss free user „tagging“ as opposed to categorization based on an a priory defined categorization schemes.

With respect to searches in the internet the early stage approach from Yahoo to perform search along given categories has been abandoned in favour of key word search as known by Google. Yet simple string matching in documents is not very precise and doesn't work at all for media files. Currently an alternative to in-advance categorization might be 'user tagging' as it is promoted most outstandingly be Shirky [12]. He refers to a service [13] that offers users to store bookmarks of web-resources and make those bookmarks available for the public. So each user who wants to remember an URL of interest can describe it with an arbitrary set of key words. Of course, each user has his own view of the resource and the description may be inaccurate or erroneous, but the assumption is, that if there are a lot of users describing the same URL, the statistics will end up establishing a widely shared set of key terms. This kind of „categorization afterwards“ lacks genuinely any hierarchy and results more in a kind of semantic net or „topic map“.

7 Discussion

There are a number of reasons why the idea of “free tagging” will not be applicable for the domain of language resources:

- The idea of „free tagging“ relies on the voluntary work of many and presupposes that the resource in question is interpretable by everybody. This is certainly not the case in the field of linguistic data, where often only the producer of the resource is able to describe it adequately.
- It is the researcher who has the deep knowledge about the construction of a corpus and about the reasons to have chosen a certain approach. This knowledge has to be stored somewhere and it’s the metadata where it is stored.
- At least the linguistic users can rely on the a priori defined categorizations, since linguistic terminology has stabilized to a large extent during the last decades.

So, tagging of the content of linguistics resources would have primarily to be done by the creator like with the rest of the metadata. On one side, the „open vocabularies“ offered currently by IMDI incite some users to slightly misuse them for an imitation of „free tagging“ e.g. if they add an overspecialized item. On the other hand “free tagging” could be an option for other “experts” to enrich the data and therefore to increase the precision and recall.

A solution and kind of promise between the two strategies may be to make every new entry „public“, e.g. adding it to the list of offered vocabulary automatically. This would benefit those who fill in the data as well as those who are querying it. Furthermore, it would inhibit users to add too specific terms by a kind of „social pressure“.

8 Conclusion

The Max-Planck-Institute Nijmegen offers several kinds of querying and browsing approaches corresponding to different user interests. The IMDI categorization scheme allows in principle for very detailed search and therefore has the potential for a high precision and high recall compared to all sorts of free text searches.

However, the IMDI forms are generally not completely filled in as was indicated in the table and even linguistic users do not fully share the same terminology. This will deteriorate the success of the searches in terms of precision and recall. Since free-text field also bear relevant information in many cases, even some linguists will prefer nevertheless a free-text search on the metadata first.

9 References

[1] http://www.mpi.nl/IMDI
[2] http://www.language-archives.org/
[3] http://www.tei-c.org/
[4] http://www.ilc.cnr.it/EAGLES96/
[5] Wittenburg, P., Peters, W., Broeder, D. (2002). *Metadata Proposals for Corpora and Lexica.* In M. Roriguez Ganzalez & C. Paz Suarez Araujo (eds.), Proceedings of the 3rd International Conference on Language Resources and Evaluation. Paris: European Language Resource Association. pp 1321-1326
[6] Broeder, D., Wittenburg, P., Crasborn, O. (2004). *Using Profiles for IMDI metadata creation.* In X. Fatima Ferreira et. al. (Eds), Proceedings of the 3rd International Conference on Language Resources and Evaluation. Paris: European Language Resource Association. pp1317-1320
[7] Aristar-Dry, H., Simons, G. (2006). *E-Meld: openness, ontologies and interoperability*. DGFS Annual Meeting on Language Documentation and Description – Working Group 6. University of Bielefeld
[8] http://lucene.apache.org/java/docs/queryparsersyntax.html
[9] http://dublincore.org/
[10] http://earth.google.com/
[11] http://www.mpi.nl/annex
[12] Shirky, Clay (2005): www.shirky.com/writings/ontology_overrated.html
[13] http://del.icio.us