IST R&D PROJECT

SHARED-COST RTD PROJECT

THEME: INFORMATION ACCESS AND INTERFACES
COMMISSION OF THE EUROPEAN COMMUNITIES
DIRECTORATE GENERAL DG INFSO

PROJECT OFFICER: DR PAT MANSON

Resource Selection and Data Fusion for Multimedia
International Digital Libraries

Identification of content and user groups

D1.2

February 4, 2003, UNIDO/WP1/Task 1.2/Version 3

Henrik Nottelmann

IST Project Number	IST-2000-26061	**Acronym**	MIND
Full title	Resource Selection and Data Fusion for Multimedia International Digital Libraries		
EU Project officer	Dr. Pat Manson		

Deliverable	**Number**	D1.2	**Name**	Identification of content and user groups		
Task	**Number**	T1.2	**Name**	Content and user groups		
Work Package	**Number**	WP1	**Name**	Architecture, Content and Context		
Date of delivery	**Contractual**		2001/03/31	**Actual**	2003/04/02	
Code name	<codename>			**Version** 3	draft ☐	final ☑
Nature	Prototype ☐ Report ☑ Specification ☐ Tool ☐ Other:					
Distribution Type	Public ☐ Restricted ☑ to: Partners, Commission, Reviewers					
Authors (Partner)	Henrik Nottelmann (UNIDO)					
Contact Person	Henrik Nottelmann					
	Email	nottelmann@ls6.cs.uni-dortmund.de	**Phone**	+49 203 379 2281	**Fax**	+49 203 379 2549
Abstract (for dissemination)	This is the revised version of D1.2 containing the current state of possible application areas, of digital libraries which could be or are covered by the MIND system and of potential user groups.					
Keywords	Application areas, Content, Digital Libraries, User groups					

Contents

Chapter 1

Content

Two possible application areas are worked out:

1. (sport) news,
2. computer science and
3. arts.

Furthermore, there are other collections that might be used for MIND.
We split these collections into three categories:

1. collections for whom a formal agreement is/may be available,
2. collections for whom it is not possible to establish a formal agreement, but which can be reused w. r. t. general copyright laws and specific license rules, and
3. collections for whom a formal agreement is not required at all.

Please note that document (an extension of version 1.2, May 18, 2001) is only a snapshot from the ongoing effort of acquiring new sources (and implementing proxies) and user groups. This document will be extended in the future.

1.1 Collections for whom a formal agreement is/may be available

1.1.1 (Sport) news

- *TREC*: text, speech

  There would be no problems in getting an agreement to use it.

  We already split up TREC-123 into 100 collections and developed small search engines and proxies for them.

  `http://trec.nist.gov`

- *Reuters*: text, images, speech

  USFD has been approved by Reuters Ltd to receive a copy of their "Corpus", comprising all of their electronically gathered news stories for the period of a year (August 1996 - 1997).

It's about 8GByte of XML-formatted news stories. USG has permission for this collection and tries to extend it for MIND.

CMU has the new Reuters data.

`http://xml.coverpages.org/ni2001-03-03-c.html`

- *BBC World Service:*: text, speech (spoken text)

  USFD has permission to use a large BBC speech data base consisting of 3,000 hours of BBC news data growing each day, it would potentially possible to gain access for the rest of the MIND partners. The caveat there is that BBC would likely refuse to allow any part of the collection to be on a public web site.

  Furthermore USDF possibly gets an archive of the BBC news online Web site, which will enable USFD to do work on this, but is not sure if USFD will be able to make this available to others though.

  `http://www.bbc.co.uk/worldservice/index.shtml`

1.1.2 Computer science

- *Daffodil*: text, facts

  Daffodil is an agent-based system combining 12 databases (e.g. ACM DL, Citeseer, DBLP) related to computer science, being developed and implemented at UNIDO (funded by the German funding organisation DFG).

  Proxies for the Daffodil wrappers (which provide access to the 12 databases in a uniform way) are under development.

  `http://www.daffodil.de`

1.1.3 Arts

- *Google:* text

  Google provides an own SOAP interface for the pure search engine, a licence is available for 1,000 queries per day.

  A proxy for Google is completed.

  `http://www.google.com`

- *Cultural Heritage Bureau for Florence, Pistoia and Prato:* images, text

  DSI reached a preliminary agreement and will get access to a collection of about 10000 images each one with catalografic (text data) info.

- *European Visual Archive*: images, textual descriptions (XML)

  USG got permission to use the data (agreed in principle).

  `http://www.cultivate-int.org/issue3/eva/`

- *London Metropolitan Archives*: images, textual descriptions (XML)

  USG got permission to use the data (agreed in principle).

  `http://www.corpoflondon.gov.uk/archives/lma/`

1.1.4 Other collections

- *Linguistic Data Consortium:* text, speech

  When the time comes, we can get other collections of data from a number of source, e.g. the Linguistic Data Consortium.

  `http://www.ldc.upenn.edu/Obtaining/`

- *Internet Movie Database:* images, text

  USG has contacted the Internet Movie Database, but is still waiting for a reply.

  `http://uk.imdb.com/`

1.2 Collections which can be reused

1.2.1 Arts

- *Web Museum:* images (searchable), facts, text annotations

  Due to the lack of a formal answer by people responsible for WebMuseum, images from this collection will be used following guidelines specified in the license agreement.

  A proxy for the Web Museum is under development.

  `http://sunsite.unc.edu/wm/`

1.2.2 Computer science

- *BIBDB:* text, facts

  We built several smaller collections from UNIDOs and USGs BIBDB (bibliographic data about computer science publications, used by bibtex) and implemented proxies for them.

1.3 Collections for whom a formal agreement is not required

1.3.1 (Sport) news

- *CNN:* text, facts, images (searchable via metadata)

  The CNN articles can be accessed via the web.

  A proxy (including parsing the HTML pages) is completed.

  `http://www.cnn.com/`

- *Times online:* text, facts

  The articles from Times Online can be accessed via the web.

  A proxy (including parsing the HTML pages) is completed.

  `http://www.timesonline.co.uk/`

- *SpeechBot:* text, facts, speech

  SpeechBot is the first internet search site for indexing streaming spoken broadcast audio from multiple sources such as On Point, PBS Online NewsHour, etc. SpeechBot does not store or serve the audio files themselves but provides users with links. Furthermore, transscripts are displayed. It's current index has over 3200 shows, 3500 hours of audio and 20 million words. The index is continually updated.

  A proxy for SpeechBot is under development.

  `http://speechbot.research.compaq.com/`

1.3.2 Arts

- *Web Gallery of Art:* text, facts, images (searchable via metadata)

  The Web Gallery of Art (providing images and metadata about paintings from multiple museums) can be accessed via the web.

  A proxy (including parsing the HTML pages) is completed.

  `http://gallery.euroweb.hu/`

- *National Gallery of Art, Washington DC:* text, facts, images (searchable via metadata)

  The National Gallery of Art, Washington DC (providing images and metadata about paintings from the museums), can be accessed via the web.

  A proxy (including parsing the HTML pages) is under development.

  `http://www.nga.giv`

- *National Gallery of Art, London:* images, facts, text annotations

  London National Gallery of Art, London(providing images and metadata about paintings from the museums), can be accessed via the web. A proxy (including parsing the HTML pages) is under development.

  `http://www.nationalgallery.org.uk/`

- *Metropolitan Museum of Art, New York:* text, facts, images (searchable via metadata)

  The Metropolitan Museum of Art, New York (providing images and metadata about paintings from the museums), can be accessed via the web.

  A proxy (including parsing the HTML pages) is under development.

  `http://www.metmuseum.org/`

- *Yahoo! Picture Gallery:* images which support text based search

  Yahoo! Picture Gallery can be used via the web for free.

  `http://gallery.yahoo.com`

1.4 Not really clear

1.4.1 Arts

- Vasari Arts Encyclopedias: images

  Status: unknown

- Giunti Multimedia Arts CDs (USG): images, text, spoken text

  Status: no reply

1.4.2 Other collections

- 100 years of National Geographic (USG): text, images

  Status: USG is in contact with the National Geographic Magazine to get access to the text and photos from the CDs.

- Alma Media (USFD): images

  Status: USFD tries to get a collection with 20,000 images from a Finish media company. The major problem is to get a answer as they are very busy.

Chapter 2

User groups

This chapter gives a brief overview of possible user groups according to the application areas.

2.1 (Sport) news

Possible user groups in the field of "(sport) news" are:

- Scottish School of Sports Studies (department at USG)
  `http://www.strath.ac.uk/sportstudies`
- Scottish School of Sports Medicine
- Hournalist's school in Sheffield

2.2 Arts

For "computer science", the following user group can be chosen:

- Computer science students in the partners departments

2.3 Arts

For "arts", the following user group can be chosen:

- Art historians in Bochum, Germany
  `http://artregister.flnet.org/`