

Programme MDCO - Edition 2007

**ANR-07-MDCO-017-01**

## **DISCO**

**Distributed Indexing and Search by Content**

### **Deliverable 3.1**

## **Description of the document collections used as test-beds**

**Version 1.1**

---

#### **Partnership**

<b>Number</b>	<b>Name</b>	<b>Type</b>
<b>1</b>	<b>Wisdom - Dauphine</b>	<b>Public laboratory</b>
<b>2</b>	<b>Réunion des Musées Nationaux (RMN)</b>	<b>EPIC</b>
<b>3</b>	<b>Wisdom - Cnam</b>	<b>Public laboratory</b>
<b>4</b>	<b>European Web Archive (EWA)</b>	<b>Association</b>
<b>5</b>	<b>IRCAM</b>	<b>Association</b>

## Document Control

---

**Title:** Description of the document collections used as test-beds

**Author:** Leïla Medjkoune (EWA), Diana Segurini (RMN)

**Editor:** EWA

**Work Package:** WP 3: Unified search by content in distributed repositories

**Status:** Version 1.1

**Date:** 17/12/2008

## Document History

---

Version	Date	Author/Editor	Description/Comments
1.0	03/12/2008	Leïla Medjkoune (EWA), Diana Segurini (RMN)	
1.1	17/12/2008	Philippe Rigaux	Slight revision

## Table of contents

---

<b>1 Introduction</b>	<b>4</b>
<b>2 Corpus Description</b>	<b>5</b>
2.1 RMN collection of images	5
2.1.1 Table of collections	5
2.1.2 Table of indexing fields	8
2.2 EWA collection of audio and video	8
2.2.1 Audio collections	10
2.2.2 Video collections	11
2.2.3 Video and audio collections	12
2.2.4 Description of items	12
<b>3 Conclusion</b>	<b>14</b>

# 1 Introduction

Disco addresses the MDCO call, Axis 2 “Algorithmes pour le traitement massif des données”. The expected results of the project are stated as follows in the Description of Work: *“a common methodology for the description and indexing of a wide range of multimedia documents, including specification and experimentation of distributed architectures addressing scalability and heterogeneity issues”*.

The goal of the WP3 3.1 deliverable is to produce a corpus of items that will be used as test-beds during the project.

As explained in the Description of Work, the whole project is content provider oriented and aims at *“dealing with the production of generic descriptors”, “producing specifications for index and search operations”, “designing distributed solution and performing evaluation”*.

In order to provide generic descriptors WP partners will define a common typology for the description of still images, videos and audio documents.

Multimedia documents archived by cultural institutions such as sound, images or videos, grow exponentially. The importance of the topic has been increased with the archiving of web resources.

Even though the number of multimedia resources collected by cultural institutions is growing, the available tools for indexing and searching are not totally fulfilling these new needs. Descriptors and indexing are remaining more suited to textual resources and to one specific institution.

Disco aims at addressing this issue by *“designing and experimenting generic and flexible techniques for content based indexing and searching, dedicated to distributed sources of multimedia documents”*.

The test-beds will therefore be used to “evaluate a multimedia document analyzer (MDA), and a distributed search engine (DSE)”. For this purpose, the corpus described in this document is containing all types of multimedia resources: images, audio and video.

## 2 Corpus Description

The corpus can be divided in three types of media:

1. Images
2. Audio documents
3. Video documents

Each type of media is provided by one content provider:

1. The collections of images are provided by the RMN, *the Réunion des Musées Nationaux*.
2. The collections of audio and video resources are provided by EWA, *the European Web Archive*.

Information provided by the partners are related to the type of resource described. As explained in WP1, images descriptions are traditionally based on high-dimensional vectors (analysis of the visual content: shape, texture, color, etc.) whereas the audio and video resources add temporal informations. The aim of Disco is to consider the multimedia resources as a whole and to extract appropriate spatial, temporal and spatiotemporal descriptors.

Under this perspective, the corpus is composed of collections which are described according to each media characteristics.

### 2.1 RMN collection of images

The corpus is composed of 8 collections containing a total of 324 993 images.

All images are either coming from the RMN personal fund (RMN photographers), or from French cultural institutions such as the Centre Pompidou or the Musée de l'Armée.

Images reproduce different types of art objects, from paintings to sculptures or manuscripts and each collection represents a specific period in history and art. This should provide an ideal basis of work for images description as most of the types of shapes, materials and techniques should be represented.

The tables contain informations such as the location of the resource, its description, its type and format for individual treatment. The number of items in each collection is also given. One of the project aim is to create and test a large scale indexing system. Collections provided as test-beds are described under this perspective in the tables below.

#### 2.1.1 Table of collections

This first table describes the collections of images provided by the RMN. The information is classified in three categories:

1. The *origin* of the collection provides informations on the institution whom owns the art objects or the collection.
2. The *description* characterizes the collection : location and type of images.
3. The last column gives the number of images composing each collection, and the total number of images composing the corpus.

Origin	Description	Number of images
RMN	Images from several museums covered by RMN photographers	237,086
RMNO	Images from the Musée d'Orsay by RMN photographers	38,210
EPMO	Images from Musée d'Orsay by the museum photographers	1,004
CNAC	Images from centre Pompidou	33,126
GUIMET	Images from Musée des arts asiatiques-Guimet located in Paris	1,045
MAP	Images from the Médiathèque de l'architecture et du Patrimoine (images donation)	4,480
SAP	Images from the Médiathèque de l'architecture et du Patrimoine (old photos collection)	817
MDA	Images from the Musée de l'Armée located in Paris	9,225
<b>Total</b>		<b>324,993</b>

The images composing these collections are homogeneous in terms of format (low definition, 600 x 800 pixels).

The type of art objects represented can change from one collection to another as well as inside the same collection. The categories of objects represented in this corpus can be listed as follows:

- Architecture
- Drawing
- Manuscript
- Art object
- Painting
- Photos
- Sculpture

This heterogeneity in the categories of object represented on the images will allow the partners to work on the description of different shapes, material and techniques.

## 2.1.2 Table of indexing fields

This second table lists the fields used for indexing the collections' images described above (tabbed CVS format).

<b>Extraction fields</b>	Inventory n°
	quotation n°
	author
	title
	description
	archive collection
	Key words
	production site
	discovery site
	period
	date
	technique/material
	dimensions (height and length)
	photographer
	summarization of quotations
	copyright
localisation	

This table contains two types of informations:

- Data related to the document localization and general description. It contains information such as the document's localization or its quotation n°.
- Data related to a more specific description of the document such as the technique used or the material represented, its dimensions, etc.

Both informations will allow DISCO partners to work on the description of images combining different types of describing attributes from the lowest to the highest level.

## 2.2 EWA collection of audio and video

The collections provided by *the European Web Archive* are composed of donations made by the Internet Archive Foundation<sup>1</sup>, located in the United States or by The National Archive of Sound and Vision in The Netherlands<sup>2</sup>.

These selected collections cover various aspects of multimedia resources, from radio shows to music recording and from TV shows to animation films. Most of these collections are in English.

40 collections are provided to the project, for an estimated total duration of 862 days and 47 hours and an approximate total number of 46,520 items.

<sup>1</sup><http://www.archive.org/index.php>

<sup>2</sup><http://portal.beeldengeluid.nl/>

Within these collections, 16 are audio collections containing music recording, live music, radio shows; 21 are video collections containing animation films, filmed lectures or interviews, TV shows as News and 3 of these collections contain audio as well as video items.

The format of the items composing the collections varies from one to another, depending on many factors as who created it and when it was created. In this regard, the collections produced by TV or radio shows should be more homogeneous than the ones offered by amateurs.

The average length of an item is about 1.5 hour. Each collection is quite homogeneous except for the free donations collections from anonymous users of The Internet Archive. These collections, as for instance "opensource\_movies", are donations of personal videos or audios and therefore the variety of format and difference in quality of items will be stronger.

Each collection has a name, a description, a duration, a total number of items and a media type. Each item is related to one collection and has a title, a location (url), a duration and a media type. In addition to these informations and because of the importance of the temporal aspect, the number of items with identified duration, the percent identified and the average length per item are specified for each collection.

The tables below describe the collections available for project use. The first three tables are organized identically. Each table describes one type of media as follows:

1. Audio collection
2. Video collection
3. Heterogeneous collections containing audio as well as video items.

The categories describing the collections are the following:

1. The name of the collection.
2. A short description of the collection which will allow partners to get a first idea of each collection content and choose for instance between a collection of live music and a collection of recorded music.
3. The total number of items composing each collection.
4. The total duration of each collection is estimated and provided in the table under this format: Days/hours/minutes/seconds (000 days 00:00:00).
5. This total duration estimation is precised with the "Items with identified duration" and the "percent identified" categories.  
"Items with identified duration" shows the number of items within one collection for which the duration has been identified.  
"Percent identified" gives the number of items for which the duration has effectively been identified per collection.
6. Based on the previous informations, the average length of one item per collection has been calculated and is reproduced as: Hours/minutes/seconds (00:00:00).
7. The audio type is specified for each collection in each table.

## 2.2.1 Audio collections

Collection Name	Description	Total Duration	Total Items	Items with identified duration	Percent identified	Average Length	Mediatype
etree	Live music shows.	582 days 19:53:15	17996	6660	37	2:06:01	Audio
opensource_audio	Audio donations from IA users and community members.	30 days 3:58:45	3751	2092	56	0:20:45	Audio
netlabels	MP3/OGG-format music of several genres.	16 days 12:22:12	2709	699	26	0:34:01	Audio
78rpm	Collection of 78rpm records and cylinder recordings released in the early 20th century.	8 days 18:17:13	673	641	95	0:19:41	Audio
naropa	Donation from the Naropa University Archive Project which is preserving and providing access to recordings made at Naropa University in Boulder, Colorado: readings, lectures, performances, seminars, panels and workshops conducted at Naropa by many of the leading figures of the U.S.literary avant-garde.	6 days 22:24:07	154	130	84	1:16:48	Audio
tse_chen_ling	Series of lectures given at the Tse Chen Ling Buddhist Center (FPMT) in San Francisco, California.	3 days 2:34:49	32	24	75	3:06:27	Audio
audio_music	Music recording of several genre.	3 days 1:36:17	189	112	59	0:39:25	Audio
presidential_recordings	Presidential speeches: public speeches made by U.S. Presidents and secret recordings made in the White House between 1940 and 1973.	2 days 21:06:37	99	86	87	0:48:12	Audio
childhood_matters	Radio shows about children care.	2 days 12:40:25	66	64	97	0:56:52	Audio
netlabel	Music recording: all styles of electronic music.	2 days 6:08:01	739	99	13	0:32:48	Audio
other_minds	Composer's Forum Panel Discussion from the 1st Other Minds Festival recorded at the Yerba Buena Center for the Arts in San Francisco.	1 day 21:54:20	145	67	46	0:41:06	Audio
audio_news	Audio news programs as Democracy Now! (news program hosted by journalists Amy Goodman and Juan Gonzalez).	13:17:16	19	15	79	0:53:09	Audio
soulseek_artists	Recording of various music genres.	5:41:08	26	15	58	0:22:44	Audio
groks	Groks Science Radio Show is a weekly science radio program heard on radio stations across the US and also as a podcast.	5:25:50	88	11	13	0:29:37	Audio
audio_misc	Collection of podcasts.	5:25:45	85	34	40	0:09:34	Audio
audio_historical	Amateur home recording of a live WWII-era radio news broadcast. Music recording (20 <sup>th</sup> Cent).	5:07:58	61	58	95	0:05:18	Audio

## 2.2.2 Video collections

Collection Name	Description	Total Duration	Total Items	Items with identified duration	Percent identified	Average Length	Mediatype
prelinger	Advertising, educational, industrial, and amateur films.	54 days 21:18:48	2974	1473	50	0:53:39	Movies
opensource_movies	Video donations from IA users and community members.	40 days 15:46:29	5544	3580	65	0:16:21	Movies
tvprograms	TV shows.	25 days 3:11:50	667	656	98	0:55:10	Movies
arsdigita	Recording of a one-year, intensive post-baccalaureate program in Computer Science based on the undergraduate course of study at the Massachusetts Institute of Technology (MIT).	8 days 23:27:17	18	12	67	17:57:16	Movies
newsandpublicaffairs	Videos or TV shows donations related to news.	3 days 20:34:40	283	202	71	0:27:29	Movies
mosaic	Television news reports from the Middle East (Egypt, Lebanon, Israel, Syria, the Palestinian Authority, Iraq and Iran).	3 days 1:14:34	71	71	100	1:01:53	Movies
open_mind	Dialogues with creative thinkers of the last half-century.	2 days 1:10:23	89	13	15	3:46:57	Movies
election_2004	Contribution from the Internet Archive users to a non-partisan collection of 2004 presidential election videos.	1 day 20:23:17	561	386	69	0:06:53	Movies
computerchronicles	Computer Chronicles was a popular television program on personal technology. It was broadcast for twenty years from 1983 – 2002.	1 day 17:27:42	765	34	4	1:13:10	Movies
arsdigitac2	Recording of University lectures (Sciences, Computer sciences, etc).	1 day 0:16:23	33	18	55	1:20:54	Movies
brick_films	"LEGO films": animation of plastic building toys, or bricks (including LEGO, Mega Bloks, Best-Lock, etc.).	21:09:40	495	252	51	0:05:02	Movies
computersandtechvideos	Videos about computer sciences, Internet (vlog, lectures, etc)	11:46:31	12	11	92	1:04:13	Movies
netcafe	Silicon Valley analyst Alex Vieux evaluates the new web startups. Shot on location at the Metreon cyber cafe in San Francisco.	9:56:21	163	9	6	1:06:15	Movies
iraq_911	A collection of archive video footage and films related to the 11 September 2001 terrorist attacks against the World Trade Center and Pentagon.	9:20:40	19	13	68	0:43:07	Movies
p2p_politics	Films (news, interviews, lectures, etc) about politics.	7:38:26	150	32	21	0:14:19	Movies
listenup	Short documentaries about society.	7:17:37	350	218	62	0:02:00	Movies
siggraph	Computer animations shown at the SIGGRAPH annual conference.	6:43:52	121	32	26	0:12:37	Movies
opensource_pets	Personal films donated by Internet Archive users (pets).	6:39:59	27	21	78	0:19:02	Movies
moviesandfilms	Documentaries and films.	6:10:46	143	22	15	0:16:51	Movies
videomisc	Films about religions (lectures, ceremonies, etc).	5:07:10	31	15	48	0:20:28	Movies

## 2.2.3 Video and audio collections

Collection Name	Description	Total Duration	Total Items	Items with identified duration	Percent identified	Average Length	Mediatype
ourmedia	Donation of works of personal media (Video blogs, photo albums, original music, documentary journalism, music videos, etc).	15 days 8:58:24	6350	4072	64	0:05:26	Movies/Audio
democracy_now	National, daily, news program airing on over 450 stations in North America.	9 days 18:48:53	425	218	51	1:04:37	Movies/Audio
conference_proceedings	A selection of speeches and lectures from tech-oriented conferences. Audio and video files.	2 days 1:33:57	107	97	91	0:30:39	Movies/Audio

## 2.2.4 Description of items

The list of items composing these multimedia collections is too large for being reproduced as such in this document. The table below shows the categories of informations supplied for each item. It was not possible to specify the duration for each of these items.

This table describes each item by giving:










1. The url where it can be accessed from.
2. The Media type of the file, either audio or video.
3. The duration of each file when identified.
4. The title of the item as given by its creator.

URL	Mediatype	Collection	Duration	Title
<a href="http://ia200106.eu.archive.org/3/audio/moe1997-01-21mdnk.dmtx.shn">http://ia200106.eu.archive.org/3/audio/moe1997-01-21mdnk.dmtx.shn</a>	audio	etree	0	moe. Live at the wetlands on 1997-01-21
<a href="http://ia200132.eu.archive.org/2/audio/moe1997-03-24dnk.shn">http://ia200132.eu.archive.org/2/audio/moe1997-03-24dnk.shn</a>	audio	etree	0	moe. Live at vogue theater on 1997-03-24
<a href="http://ia200134.eu.archive.org/1/audio/nero2003-06-25">http://ia200134.eu.archive.org/1/audio/nero2003-06-25</a>	audio	etree	0	nero Live at B-Side on 2003-06-25
<a href="http://ia200128.eu.archive.org/3/audio/nero2002-04-26.matrix.shnf">http://ia200128.eu.archive.org/3/audio/nero2002-04-26.matrix.shnf</a>	audio	etree	0	nero Live at The Comfort Zone on 2002-04-26
<a href="http://ia200136.eu.archive.org/1/audio/nero2002-12-20.matrix.shnf">http://ia200136.eu.archive.org/1/audio/nero2002-12-20.matrix.shnf</a>	audio	etree	0	nero Live at The Rivoli on 2002-12-20
<a href="http://ia200134.eu.archive.org/2/audio/nero2003-11-09.shnf">http://ia200134.eu.archive.org/2/audio/nero2003-11-09.shnf</a>	audio	etree	0	nero Live at Toad's Place on 2003-11-09
<a href="http://ia200116.eu.archive.org/0/audio/Rugburns1990-10-19.shnf">http://ia200116.eu.archive.org/0/audio/Rugburns1990-10-19.shnf</a>	audio	etree	0	rugburns Live at The Blarneystone on 1990-10-19
<a href="http://ia200042.eu.archive.org/2/items/ch2006-02-23.flac16">http://ia200042.eu.archive.org/2/items/ch2006-02-23.flac16</a>	audio	etree	10002	Charlie Hunter Live at The Haunt on 2006-02-23
<a href="http://ia200111.eu.archive.org/0/audio/sci1997-02-18.sbd.shnf">http://ia200111.eu.archive.org/0/audio/sci1997-02-18.sbd.shnf</a>	audio	etree	10002	String Cheese Incident Live at The Mangy Moose on 1997-02-18
<a href="http://ia201130.eu.archive.org/3/audio/um2005-04-15.flac16">http://ia201130.eu.archive.org/3/audio/um2005-04-15.flac16</a>	audio	etree	10002	Umphey's McGee Live at Canopy Club on 2005-04-15

An item can be described as a sub directory containing a number of files from two types: the audio or video file itself (e.g: .mov) and the files containing metadata (e.g.: .xml).

The example below shows the type of information available when accessing an item by the url given in the above table:

The item "Charlie Hunter Live at The Haunt on 2006-02-23" from *etree* audio collection, accessible at <http://ia200042.eu.archive.org/2/items/ch2006-02-23.flac16>

<b>Index of /2/items/ch2006-02-23.flac16</b>				
<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>	
 <a href="#">Parent Directory</a>	28-Feb-2006 20:28	-		
 <a href="#">ch2006-02-23.flac16.ffp</a>	27-Feb-2006 12:49	1k		
 <a href="#">ch2006-02-23.flac16 64kb.m3u</a>	27-Feb-2006 14:21	1k		
 <a href="#">ch2006-02-23.flac16 files.xml</a>	27-Feb-2006 14:21	31k		
 <a href="#">ch2006-02-23.flac16 meta.xml</a>	27-Feb-2006 23:08	2k		
 <a href="#">ch2006-02-23.flac16 vbr.m3u</a>	27-Feb-2006 14:21	1k		
 <a href="#">ch2006-02-23.txt</a>	26-Feb-2006 22:39	1k		
 <a href="#">disc1/</a>	27-Feb-2006 14:10	-		
 <a href="#">disc2/</a>	27-Feb-2006 14:13	-		

*Apache/1.3.33 Server at ia200042.eu.archive.org Port 80*

### 3 Conclusion

The aim of the D3.1 document is to provide a description of the selected corpus that will be used as test-beds during the project.

For description purpose, the corpus has been divided in two parts related to the content provider of collections as well as the type of media.

The *Réunion des Musées Nationaux* provides the collections of images and *the European Web Archive*, the collections of audio and video resources.

Each of these collections has been described according to the media specificities from the perspective of indexing and search applications development.

Images descriptions emphasize on spatial informations whereas audio and video require temporal rather than static informations.

The final corpus should allow Disco partners to proceed with the development and evaluations throughout the project, respecting every of the needs expressed:

- An heterogeneous corpus of images, audio and video items to study and develop descriptors automatic production.
- Multidimensional index structures (indexing tool application).
- A wide range of collections to test scalability and heterogeneity in a distributed environment.

The content and environment provided by RMN and EWA should allow the implementation of a platform to accomplish tests on real data and real environments.