D1.2: Unpublished data

Unpublished data is spread over different locations:
- C4DM_datasets server
- C4DM_scratch
- local machines

Reasons for not publishing the data:
- copyrights: it is usually shared internally on the servers
- published elsewhere: the data was retrieved from other repositories and made available internally on C4DM servers
- raw data that needs to be prepared for publication
- no interest in publishing it

This is a list of the data currently available on the C4DM servers (C4DM_datasets), which does not have public access. If an external source is given, the data set has been already published elsewhere. More information can be found in the file "C4DM_data_servers.xml"

Name	***Type***	***Source***	***Legal***
Auditory lab sound events	audio, video	www.auditorylab.org	?
Birdsong recordings	audio	mixed (CD, internal, external)	partly distributable
Breakbeats collection	audio	Various	Probably copyrighted
C4DM Multritrack Collection	audio, metadata	Mixed	mixed (folders structure reflects copyrights)
C4DM Music Collection	audio	Various (CDs)	Mostly copyrighted
IViE speech dataset	audio	www.phon.ox.ac.uk/files/apps/IViE/	?
Magnatune collection	audio	www.magnatune.com	CC
Marine bioacoustics collection	audio, other	external	?
Mathieu Thomas dereverb	.rar (audio)	?	?
Midi collection	midi, text, inf	www.geerdes.de	copyrighted
MillionSongDataset	analysis and	labrosa.ee.columbia.edu/millionsong/	GNU?

	metadata		
MIREX 2006 evaluation datasets	audio, analysis data	www.music-ir.org/evaluation/MIREX/data/2006/	?
olpc-sound-samples	audio, metadata	wiki.laptop.org/go/Sound_samples	CC
RWC collection	audio	staff.aist.go.jp/m.goto/RWC-MDB/	Copyrighted
RWC Annotations	annotations, midi	internal?	?
RWC Ground-Truth	annotations , midi	staff.aist.go.jp/m.goto/RWC-MDB/	?
TempermentProject	audio, annotations, midi	internal, external	?
Papers TestSets	audio, annotations, midi	internal, external	Various (see specific paper)
TIMIT	audio, annotations, code	CD?	Probably copyrighted
Unsorted data	?	?	?

On top of the data listed on the server, more data is available on local machine owned by the single researchers at C4DM. To audit this data we interviewed some of our colleagues, and sent out an online survey. More interviews are programmed. This is an incomplete list of their unpublished data:

Name	***Type***	***Copyrights***
Small set of labeled onsets in voice tracks	Audacity labels	CC
Onset labels on multitrack data	Audio, RDF, SonicVisualizer	CC, other
SoundOnSound web scraping data	RDF?	Copyrighted
Hotttabs data	Youtube videos, chord transcr.	Copyrighted
Youtube data analysis database	Mysql database	Copyrighted
Ethnographic study notebooks	Paper	Sensitive data
Auditory signals	Similar to audio	No

Auditory model parameters	Text files	No
Listening test responses	Text files	No
Impossible Alone evaluation files	Video, sound, software logs, transcribed interviews, code	Sensitive data
Study on effects of constraints on interactive music systems	Scanned questionnaires, video, software logs, notes	Confidential data
Test recordings	Audio and video	Maybe
Yin and SWIPE pitch estimates for standard datasets	Matlab data files	Parts
Results of performance tests on optimisation algorithms	Matlab data files	No
Estimation of audio pitch trajectories using MIDI	Matlab files, MIDI, WAV audio	Parts
Anatomy of (music) charts	Text files	Probably (last.fm)
Mixtures of instrument phrases	Wav, midi	Parts
Environmental sounds data	Wav	No
Harpsichord temperament estimation recordings (see also C4DM-datasets server)	Wav	No
LinkedBrainz RDF dump	RDF	No