# FTPMirror

Dr. Andrew C.R. Martin

May 19, 2014

## 1 Introduction

FTPMirror is a simple program to mirror one or more FTP sites and to handle compression or decompression of the files. The Sunsite Mirror script (and indeed the Perl LWP package) seems to fail on very large remote directories (e.g. the PDB which has > 67700 files). This script will handle such large directories.

FTPMirror uses a single configuration file for all the sites to be mirrored. It can mirror single files or directories and can handle directory trees if required.

## 2 Installation

Note that you must have the following software installed to use FTPMirror:

| | |
|---|---|
| Perl | the Perl interpreter |
| Perl LWP | the LWP package for Perl (plus any pre-requisites that it may need) |
| wget | The wget program for downloading files by FTP or HTTP. Only required if you are mirroring very large directories using wget mode. |

Once you have downloaded the FTPMirror Perl script, simply place it somewhere in your path and ensure it is executable:

```
chmod +x ftpmirror.pl
```

## 3 Running FTPMirror

FTPMirror is run simply by typing the command:

```
ftpmirror.pl configfile
```

where configfile is a configuration file as described below.

You can run the program with a `-h` flag to obtain help:

```
ftpmirror.pl -h
```

and with debugging options:

```
ftpmirror.pl -debug=n configfile
```

The `-debug` flag may be set to values 1, 2, or 3 with the following effects:

1. Prints results of parsing the configuration file, wget reports progress
2. Mirroring exits after 10 files

3. When cleaning local files, the first 10 files from the remote file and directory lists are printed

You can also run the program with a `-quiet` flag to suppress all output about progress of what files are being downloaded or removed.

As of V1.4 (19.05.14), the default behaviour is not to delete any local files if more than 50% of files have gone away on the remote compared with the current local copy. This is designed to prevent problems with the connection dropping while getting the remote directory listing. The `-forcedelete` option overrides this behaviour for all mirrors and deletes the files anyway. You can also use `forcedelete` option in the config file for an individual mirror if you know it is very dynamic and often deletes a lot of files.

# 4 The configuration file

The configuration file consists of lines with two compulsory fields (the source and destination) and optional fields containing flags as described below. **All fields must appear on a single line.**

The configuration file may contain blank lines and comments introduced with a hash character(#).

The first field is the source URL in the form:

```
ftp://server/directory/
```

For example:

```
ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/
```

The second field contains the full path to the destination directory in which the mirror will be stored:

```
/path/to/mirror/
```

For example:

```
/data/pdb/
```

This may also be a filename (rather than a directory name) if the source URL is a file rather than a directory and the 'file' flag is used (see below).

Remaining fields are flags:

**recurse** recurse into lower directories. By default only the current directory specified by the URL will be mirrored. Use this option to mirror sub-directories as well.

**decompress** decompress remote compressed files. Currently this only works with gzipped remote files.

**compress** compress remote uncompressed files. Currently this only uses gzip to compress the local files.

**file** this specifies that the remote URL refers to a single file rather than a directory

**wget** use wget rather than LWP to obtain the remote directory listing. This is used for big directories where LWP seems to fail.

**retry=n** this is only valid when 'wget' is used and specifies the number of wget retries (Default is 1). A value of 0 will keep trying indefinitely.

**noclean** do not clean up local files that have gone away on the remote machine

**forcedelete** As of V1.4 (19.05.14), the default behaviour is not to delete any local files if more than 50% of files have gone away on the remote compared with the current local copy. This is designed to prevent problems with the connection dropping while getting the remote directory listing. The forcedelete option overrides this behaviour and deletes the files anyway. You can also use `-forcedelete` on the command line to achieve the same thing for all mirrors.

**fast** just checks if files have appeared/disappeared rather than checking the datestamps on the files. This means that if the content of a file has changed, that will not be reflected in the mirror. The default is to check the files more carefully by comparing date stamps.

**regex=r** only downloads files if the filename (excluding the path) matches the specified Perl regular expression. This is a full Perl regular expression, so may contain alternatives, anchors, etc.

**excl=r** skip files if the full URL filename path matches the specified Perl regular expression. This is a full Perl regular expression, so may contain alternatives, anchors, etc. This overrides matching with regex= thus allowing you to retrieve all files matching a certain pattern unless they are in directories which match another pattern

# 5 Examples

The following examples all represent individual lines in the configuration file. **Each should be entered on a single line in the config file, even if broken across multiple lines here.**

Mirror the PDB:

```
ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/ /acrm/data/pdb/ wget
```

This would be slow since it is comparing the date stamps on each file.

Mirror the PDB, but don't check the date stamps to speed things up:

```
ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/ /acrm/data/pdb/ wget fast
```

Mirror the PDB, decompressing each local file:

```
ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/ /acrm/data/pdb/ wget fast
      decompress
```

Mirror the Human SubXXXX.bcp.gz files from dbSNP

```
ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/database/organism_data/
      /tmp/mirror/ regex=^Sub.*\.bcp\.gz
```

Mirror the Human SubXXXX.bcp.gz files from dbSNP but exclude those that are in archive directories

```
ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/ /tmp/mirror/
      regex=^Sub.*\.bcp\.gz excl=archive recurse
```