Setting up a Data Source

Guidelines and Conventions

TR-MRCH-DCC-GY005

The PhenoDCC Team

developers@har.mrc.ac.uk

Dated 10 October 2014

Overall process

Various research centres carry out phenotyping by running several experiments. These experiments generate large amounts of data. For further analysis of this data with data generated by other centres, the *data coordination centre* (DCC) collects and organises them at a central data repository using a common data exchange and storage format.

The experimental results are captured using centre-specific *LIMS*. Although there is no restriction on the usage of a specific LIMS, all of the data must be exported as *XML* documents using the IMPC data-exchange format defined in the DCC *XSD* specification and IMPReSS.

To reduce data transfer time, all of the XML documents are required to be made available as a compressed data archive file using the *Zip* file format.

Once the compressed files are ready for dissemination, they are copied to the centre-specific data source (either *FTP* or *SFTP* server) that was agreed with the DCC. Finally, in the next *crawling* session at the DCC, all of the new files currently available at the data sources will be downloaded and processed.

If there are any issues with the data, these will be captured in the *PhenoDCC Tracker* web application.

* Important topics, concepts or terms are highlighted using blue italicised text. Texts on the right-hand margin provide details on terms and abbreviations.

LIMS
A Laboratory Information Management System (LIMS) helps scientists record the setup and results of an experiment.

XSD and XML
Extensible Markup Language (XML) allows structured and typed data exchange. The XML Schema Definition (XSD) language is used to specify the structure and data types for a valid XML document.

PhenoDCC Crawler
A system that periodically checks the data sources, and retrieves and processes data files that have not already been processed.

PhenoDCC Tracker
A system that captures the provenance, status and contents of a data archive as it was downloaded and processed by the PhenoDCC crawler.

Data archive files and their contents

- A data archive file is allowed to hold several files.
- Not all of these files are expected to be XML documents with phenotype data. For instance, it is possible to include logging information files etc.
- Only XML documents that match the XML document naming convention are processed by the crawler.
- All XML documents that are to be processed *must exist at the root of the archive*. In other words, they should not reside in sub-directories inside the archive file. For instance, the following is a valid data archive (highlighted directory and files will be ignored by the crawler):

```
H.2013-03-05.1.impc.zip
     |____ H.2013-03-05.1.specimen.impc.xml
     |____ H.2013-03-05.2.specimen.impc.xml
     |____ H.2013-03-05.1.experiment.impc.xml
     |____ H.2013-03-05.2.experiment.impc.xml
     |____ info
         |____ log.xml
         |____ README.txt
```

- Each of the *data archive files are processed independently*. Hence, two data archive files can contain files that share the same file names. For instance, the following data archive files are valid:

```
H.2013-03-05.1.impc.zip
|____ H.2013-03-05.1.specimen.impc.xml
|____ H.2013-03-05.2.specimen.impc.xml
```

```
H.2013-03-05.2.impc.zip
|____ H.2013-03-05.1.specimen.impc.xml
|____ H.2013-03-05.2.specimen.impc.xml
```

Data archive file naming convention

All data archive files must use the following naming convention:

```
[a-zA-Z_]*[.][0-9]{4}-[0-9]{2}-[0-9]{2}[.][0-9]{1,5}[.]impc[.]zip
```

The tokens are:

<ILAR code>.<Year>-<Month>-<Day>.<Increment>.impc.zip

For instance, the following are valid filenames for data archive files:

```
H.2013-03-05.1.impc.zip
WTSI.2013-03-05.2.impc.zip
```

Increment number

For each centre, all of the data archive files at a given data source must have a *unique name*. This is how the crawler tracks if the file has already been processed. Hence, if a file is to be resubmitted* with different contents (say, with fixes etc.) please update the *increment number*. For instance,

```
H.2013-03-05.1.impc.zip → H.2013-03-05.2.impc.zip
WTSI.2013-03-05.2.impc.zip → WTSI.2013-03-05.3.impc.zip
```

* Please note that simply deleting and replacing the file will not work. The crawler maintains a history of all the files that it has already processed in previous sessions. Hence, increment numbers serve as version numbers.

XML document naming convention

All XML documents must use the following naming convention:

```
[a-zA-Z_]*[.][0-9]{4}-[0-9]{2}-[0-9]{2}[.][0-9]{1,5}[.](specimen|experiment)[.]impc[.]xml
```

The tokens are:

<ILAR code>.<Year>-<Month>-<Day>.<Increment>.(specimen | experiment).impc.xml

For instance, the following are valid filenames for XML documents

```
H.2013-03-05.1.experiment.impc.xml
H.2013-03-05.1.specimen.impc.xml
```

Increment number

All of the files inside the data archive file must have a *unique name*. This is how the crawler tracks the processing status of XML documents within a data archive. There are cases when the phenotype data is best split into multiple XML documents. In these cases, we use the *increment number* to make a group of related XML documents have unique names while sharing common identifiers. For instance,

```
H.2013-03-05.01.impc.zip
   |____ H.2013-03-05.1.specimen.impc.xml
   |____ H.2013-03-05.2.specimen.impc.xml
   |____ H.2013-03-05.1.experiment.impc.xml
   |____ H.2013-03-05.2.experiment.impc.xml
```

Configuring the data source

- The data source server must be either a FTP, or a SFTP server.
- It must be made accessible from the outside world.
- A user account must be created specifically for the crawler to use, preferably **`dcccrawler`**.
- Write permissions must be disabled for this user account.
- The IMPC directory for this user must contain three sub-directories (directory names are case-sensitive):

  add – Put all of the data archive files inside this directory.
  delete – *not used at the moment*
  edit – *not used at the moment*

The following is an example setup:

```
Hostname: sftp.example-centre.ac.uk, IMPC path: /home/dcccrawler/some/path/to/IMPC
Username: dcccrawler, Password: <some password>

$ cd /home/dcccrawler/some/path/to/IMPC
$ tree
.
|__add
|    |__H.2013-03-05.1.impc.zip
|    |__H.2013-05-05.1.impc.zip
|__delete
|__edit
```

Requirements for PhenoDCC Crawler

- Hostname or IP address for FTP, or SFTP server
- Username and password
- IMPC path *if required*; otherwise, crawler assumes that base path is user's home directory

The following are example setups:

```
Hostname: ftp.example-centre.ac.uk, IMPC path: <undefined>
Username: impc_user, Password: <some password>

Crawler will process /home/impc_user/add
```

```
Hostname: sftp.example-centre.ac.uk, IMPC path: /home/dcccrawler/some/path/to/IMPC
Username: dcccrawler, Password: <some password>

Crawler will process /home/dcccrawler/some/path/to/IMPC/add
```

* Please ensure that all of the files in these directories have read permissions enabled, so that they can be opened and downloaded by the PhenoDCC Crawler.

Uploading media files

To upload media files to the PhenoDCC, the procedure is as follows:

- In the XML document submitted by the centre, the centre must specify the full URI of the media files as the values of the parameters they wish to submit. For instance, in the following example XML document, a centre is submitting media files for IMPC_XRY_034_001, IMPC_XRY_048_001 and IMPC_XRY_050_001.

- The centre should then make the media files accessible externally by putting the relevant files on the HTTP/FTP/sFTP servers chosen by the centre. For instance, to make the example XML document complete, the centre should put the following files in their FTP server `ftp://ftp.example.org`.

```
/home/images/Xray/file1.dcm
/home/images/Xray/file2.dcm
/home/images/Xray/file3.dcm
```

It is possible to use the same FTP/sFTP server where the Zip files are made available. It is advisable, however, to put these inside a separate media files directory, and not forget to specify the full path in the XML documents. The media downloader at the PhenoDCC will use the same authentication credentials as used to download the Zip files.

After the PhenoDCC has processed the XML document, it will automatically download the media files specified in the XML document and carry out the necessary processing.

Uploading media files

Here is an example XML document with media files:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<centreProcedureSet xmlns="http://www.mousephenotype.org/dcc/exportlibrary/datastructure/core/procedure"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocat\
ion="http://www.mousephenotype.org/dcc/exportlibrary/datastructure/core/procedure
http://www.mousephenotype.org/dcc/exportlibrary/datastructure/core/procedure/procedure_defin\
ition.xsd">
<centre centreID="Centre" pipeline="IMPC_001" project="IMPC">
  <experiment experimentID="Xray_1679" dateOfExperiment="2014-10-01">
    <specimenID>AAWU_60_171260</specimenID>
    <procedure procedureID="IMPC_XRY_001">
      ...
      <seriesMediaParameter parameterID="IMPC_XRY_034_001">
        <value incrementValue="1" URI="ftp://ftp.example.org/home/images/Xray/file1.dcm" fileType="img/dicom"/>
      </seriesMediaParameter>
      <seriesMediaParameter parameterID="IMPC_XRY_048_001">
        <value incrementValue="1" URI="ftp://ftp.example.org/home/images/Xray/file2.dcm" fileType="img/dicom"/>
      </seriesMediaParameter>
      <seriesMediaParameter parameterID="IMPC_XRY_050_001">
        <value incrementValue="1" URI="ftp://ftp.example.org/home/images/Xray/file3.dcm" fileType="img/dicom"/>
      </seriesMediaParameter>
      ...
    </procedure>
  </experiment>
</centre>
</centreProcedureSet>
```

Caveats and trouble-shooting

- Please ensure that all of the files and directories have read permissions enabled. Without this, the PhenoDCC Crawler will be unable to download the files.

- Please note that, at the moment, tokens in the data archive file names and XML document file names are treated as strings; hence, the PhenoDCC crawler treats the following files as different:

  `H.2013-03-05.1.experiment.impc.xml` and
  `H.2013-03-05.00001.experiment.impc.xml`.

  However, in the future, the PhenoDCC Crawler might start using the token values.

- Furthermore, as a consequence of the above point, it does not matter at the moment how the increment number changes relative to existing files. For instance, the following change to the increment number is currently allowed:

  `H.2013-03-05.1.experiment.impc.xml →`
  `H.2013-03-05.10.experiment.impc.xml`.

  Nevertheless, we would advise against such changes as it might confuse future updates of the PhenoDCC Crawler, when it starts making sense of the file name tokens.