Infrastructure - Task #3076

Task # 3635 (New): Phase two Morpho implementation

Allow user to specify Morpho storage directory

2012-07-17 00:09 - Jing Tao

Status:	New	**Start date:**	2012-07-16
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Morpho	**Estimated time:**	0.00 hour
Target version:	2013.10-Block.2.1		
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

Currently, morpho's data file structure looks like:

/user-home-directory/.morpho/profiles/profile-name/data/scope-name/12.2
/cache/scope-name/12.1
/temp/scope-name/1.3
/incomplete/scope-name/2.2

The combination of the scope-name and file name is the identifier. Since new morpho will support different identifier system, such as DOI and UUID, the current data file system wouldn't work.
Moreover, some morpho users complained that the directory which stores the data is opaque.

Here is the proposed new file structure:
/user-home-directory/.morpho/profiles/profile-name/cache
/cache/sysmeta
/temp
/temp/sysmeta
/incomplete
/incomplete/sysmeta

The data directory will be separated from .morpho directory. Users can specify any directory which he/she wants.

The first step we will create a new property name "morphoDataDirectory" in the configuration file. The value will be the data directory path. The default value is "MorphoData". If it is the default value, the file path will be /user-home-directory/MorphoData. Under the specified the data directory, morpho will automatically generates a sysmeta directory to store the system metadata if the directory doesn't exist.

The second step, morpho will provide a GUI for users to specify the directory. Also, morpho may support multiple data directories.

History

#1 - 2012-07-17 16:43 - Jing Tao

Matt also suggest to separate out the meta, ore and data files from data directory into different subdirectories in order to overcome potential file system limits on number of files.

#2 - 2012-07-29 22:56 - Jing Tao

Structure in the .morpho:
/user-home-directory/.morpho/profiles/profile-name/cache/data
/user-home-directory/.morpho/profiles/profile-name/cache/data/sysmeta
/user-home-directory/.morpho/profiles/profile-name/cache/metadata
/user-home-directory/.morpho/profiles/profile-name/cache/metadata/sysmeta
/user-home-directory/.morpho/profiles/profile-name/cache/ore
/user-home-directory/.morpho/profiles/profile-name/cache/ore/sysmeta

Traditionally, morpho keeps the unsaved data file in temp. However, if user clear the temp directory even though he may have incomplete package. This can cause the package miss data files. So I propose the unsaved data file will be stored in incomplete directory which can be persistent.

/user-home-directory/.morpho/profiles/profile-name/incomplete/data
/user-home-directory/.morpho/profiles/profile-name/incomplete/data/sysmeta
/user-home-directory/.morpho/profiles/profile-name/incomplete/metadata
/user-home-directory/.morpho/profiles/profile-name/incomplete/metadata/sysmeta
/user-home-directory/.morpho/profiles/profile-name/incomplete/ore
/user-home-directory/.morpho/profiles/profile-name/incomplete/ore/sysmeta

Some other files wil be stored in temp (i am not sure if we will need it any more).
/user-home-directory/.morpho/profiles/profile-name/temp/

After user specifies a local store, the directory structure will look like:
/user-specified-directory/data
/user-specified-directory/data/sysmeta
/user-specified-directory/metadata
/user-specified-directory/metadata/sysmeta
/user-specified-directory/ore
/user-specified-directory/ore/sysmeta

#3 - 2012-07-29 23:00 - Jing Tao

I am not sure if the user specified data store will be base on the profile or whole morpho user. But i prefer to base on each profile.

#4 - 2012-07-31 18:43 - Chris Jones

In order to allow data managers to have more control over file names for their data files and metadata files, morpho needs a means of tracking local filenames against the object identiifier, and optionally tracking a 'network copy' of the object. For instance, morpho could maintain an internal h2 SQL database with a table such as:

pid	local_uri	dataone_uri
doi:10.6085/AA/CMRX00_XXXITBDXLSR02_20060618.50.5	file:///Users/frenockm/PISCO/metadata/CMRX00_XXXITBDXLSR02_20060618.50.5.xml	https://cn.dataone.org/cn/v1/resolve/doi:10.6085/AA/CMRX00_XXXITBDXLSR02_20060618.50.5
...		

This may also be implemented as a SOLR index as opposed to SQL tables. When the user changes a local URI outsie of Morpho (renames a file), Morpho may need to bring up a dialog asking for the location of the missing file (likely on startup).

#5 - 2012-08-01 23:39 - Jing Tao

Chris' suggestion is great. We will address the issue on the task 3074.

The user specified directory only store the eml documets generating by morpho, and the downloaded eml and data files from network by Morpho. The local file which is imported to morpho will keep at the original location. Morpho wouldn't copy it to its local store, just assign it an identifier and the identifer-filename.mapping will redirect the identifier to the local file location. If user modify the local file by an external editor, then the next time morpho open the package and it will prompt the user either to assign a new id to the local file or ask user to import the data file again depending on the nature of the change. If user modify the file in morpho, morpho will copy the old version to the central store with the hash string of the old identifier and overwrite the orginal file. Of course a new identifier will be assigned to the new file.

Open issue:
Where we store the system metadata and ore information?

1. Store them in the file system described above.
   advantage - they persist in a safer way than in an embedded database. It is easy to take a look for the trouble shooting.
   disadvantage - every time we need to parse them to get the information. But those files are pretty small, it maybe is not very expensive.

2. Store them in a embedded database.
   advantage - we don't need to parse the documents and only do sql commands to get information.
   disadvantage - if the database crashes, it will be a disaster. It is hard to take a look for the trouble shooting since it is a binary format.

#6 - 2012-10-11 15:33 - Dave Vieglais

- Target version changed from Sprint-2012.37-Block.5.3 to Sprint-2012.41-Block.6.1

#7 - 2012-10-14 15:00 - Ben Leinfelder

- Subject changed from Morpho's new data file system. to Allow user to specify Morpho storage directory

- Category set to Morpho

Moving this to 2.x.y target. I think the structure of the filesystem looks like a fine idea, but allowing users to directly edit these files opens up a huge cn of worms in terms of consistency and versioning.

#8 - 2012-10-24 18:20 - Ben Leinfelder

- Target version changed from Sprint-2012.41-Block.6.1 to Sprint-2012.44-Block.6.2

#9 - 2012-12-12 16:51 - Chris Jones

- Target version changed from Sprint-2012.44-Block.6.2 to Sprint-2012.50-Block.6.4

#10 - 2013-03-01 18:33 - Ben Leinfelder

- Target version changed from Sprint-2012.50-Block.6.4 to 2013.10-Block.2.1

#11 - 2013-03-02 05:29 - Ben Leinfelder

- Parent task changed from #3075 to #3635