NAME

ModifySDFilesDataFields.pl - Modify data fields in SDFile(s)

SYNOPSIS

ModifySDFilesDataFields.pl SDFile(s)...

ModifySDFilesDataFields.pl [-d, --detail infolevel] [--datafieldscommon newfieldlabel, newfieldvalue, [newfieldlabel, newfieldvalue,...]] [--datafieldsmap newfieldlabel, oldfieldlabel, [oldfieldlabel,...]; [newfieldlabel, oldfieldlabel, [oldfieldlabel,...]]] [--datafieldsmapfile filename] [--datafieldURL URLDataFieldLabel, CGIScriptPath, CGIParamName, CmpdIDFieldLabel] [-h, --help] [-k, --keepolddatafields all | unmappedonly | none] [-m, --mode molname | datafields | both] [--molnamemode datafield | labelprefix] [--molname datafieldname or prefixstring] [--molnamereplace always | empty] [-o, --overwrite] [-r, --root rootname] [-w, --workingdir dirname] SDFile(s)...

DESCRIPTION

Modify molname line and data fields in *SDFile(s)*. Molname line can be replaced by a data field value or assigned a sequential ID prefixed with a specific string. For data fields and modification of their values, these types of options are supported: replace data field labels by another set of labels; combine values of multiple data fields and assign a new label; add specific set of data field labels and values to all compound records; and others.

The file names are separated by space.The valid file extensions are *.sdf* and *.sd*. All other file names are ignored. All the SD files in a current directory can be specified either by **.sdf* or the current directory name.

OPTIONS

-d, --detail *infolevel*

Level of information to print about compound records being ignored. Default: *1*. Possible values: *1, 2 or 3*.

--datafieldscommon *newfieldlabel, newfieldvalue, [newfieldlabel, newfieldvalue,...]*

Specify data field labels and values for addition to each compound record. It's a comma delimited list of data field label and values pair. Default: *none*.

Examples:

```
DepositionDate,YYYY-MM-DD
Source,www.domainname.org,ReleaseData,YYYY-MM-DD
```

--datafieldsmap *newfieldlabel, oldfieldlabel, [oldfieldlabel,...]; [newfieldlabel, oldfieldlabel, [oldfieldlabel,...]]*

Specify how various data field labels and values are combined to generate a new data field labels and their values. All the comma delimited data fields, with in a semicolon delimited set, are mapped to the first new data field label along with the data field values joined via new line character. Default: *none*.

Examples:

```
Synonym,Name,SystematicName,Synonym;CmpdID,Extreg
HBondDonors,SumNHOH
```

--datafieldsmapfile *filename*

Filename containing mapping of data fields. Format of data fields line in this file corresponds to --datafieldsmap option. Example:

```
Line 1: Synonym,Name,SystematicName,Synonym;CmpdID,Extreg
Line 2: HBondDonors,SumNHOH
```

--datafieldURL *URLDataFieldLabel, CGIScriptPath, CGIParamName, CmpdIDFieldLabel*

Specify how to generate a URL for retrieving compound data from a web server and add it to each compound record. *URLDataFieldLabel* is used as the data field label for URL value which is created by combining *CGIScriptPath,CGIParamName,CmpdIDFieldLabel* values: CGIScriptPath?CGIParamName=CmpdIDFieldLabelValue. Default: *none*.

Example:

```
Source,http://www.yourdomain.org/GetCmpd.pl,Reg_ID,Mol_ID
```

-h, --help

Print this help message.

-k, --keepolddatafields *all* | *unmappedonly* | *none*

Specify how to transfer old data fields from input SDFile(s) to new SDFile(s) during *datafields* | *both* value of -m, --mode option: keep all old data fields; write out the ones not mapped to new fields as specified by --datafieldsmap or <--datafieldsmapfile> options; or ignore all old data field labels. For *molname* -m --mode, old datafields are always kept. Possible values: *all* | *unmappedonly* | *none*. Default: *none*.

-m, --mode *molname* | *datafields* | *both*

Specify how to modify SDFile(s): *molname* - change molname line by another datafield or value; *datafield* - modify data field labels and values by replacing one label by another, combining multiple data field labels and values, adding specific set of data field labels and values to all compound, or inserting an URL for compound retrieval to each record; *both* - change molname line and datafields simultaneously. Possible values: *molname* | *datafields* | *both*. Default: *molname*

--molnamemode *datafield* | *labelprefix*

Specify how to change molname line for -m --mode option values of *molname* | *both*: use a datafield label value or assign a sequential ID prefixed with *labelprefix*. Possible values: *datafield* | *labelprefix*. Default: *labelprefix*.

--molname *datafieldname or prefixstring*

Molname generation method. For *datafield* value of --molnamemode option, it corresponds to datafield label name whose value is used for molname; otherwise, it's a prefix string used for generating compound IDs like labelprefixstring<Number>. Default value, *Cmpd*, generates compound IDs like Cmpd<Number> for molname.

--molnamereplace *always* | *empty*

Specify when to replace molname line for -m --mode option values of *molname* | *both*: always replace the molname line using --molname option or only when it's empty. Possible values: *always* | *empty*. Default: *empty*.

-o, --overwrite

Overwrite existing files.

-r, --root *rootname*

New SD file name is generated using the root: <Root>.<Ext>. Default new file name: <InitialSDFileName>ModifiedDataFields.<Ext>. This option is ignored for multiple input files.

-w, --workingdir *dirname*

Location of working directory. Default: current directory.

EXAMPLES

To replace empty molname lines by Cmpd<CmpdNumber> and generate a new SD file NewSample1.sdf, type:

```
% ModifySDFilesDataFields.pl -o -r NewSample1 Sample1.sdf
```

To replace all molname lines by Mol_ID data field generate a new SD file NewSample1.sdf, type:

```
% ModifySDFilesDataFields.pl --molnamemode datafield
--molnamereplace always -r NewSample1 -o Sample1.sdf
```

To replace all molname lines by Mol_ID data field, map Name and CompoundName to a new datafield Synonym, and generate a new SD file NewSample1.sdf, type:

```
% ModifySDFilesDataFields.pl --molnamemode datafield
  --molnamereplace always --molname Mol_ID --mode both
  --datafieldsmap "Synonym,Name,CompoundName" -r
  NewSample1 -o Sample1.sdf
```

To replace all molname lines by Mol_ID data field, map Name and CompoundName to a new datafield Synonym, add common fields ReleaseDate and Source, and generate a new SD file NewSample1.sdf without keeping any old SD data fields, type:

```
% ModifySDFilesDataFields.pl --molnamemode datafield
  --molnamereplace always --molname Mol_ID --mode both
  --datafieldsmap "Synonym,Name,CompoundName"
  --datafieldscommon "ReleaseDate,yyyy-mm-dd,Source,
  www.mayachemtools.org" --keepolddatafields none -r
  NewSample1 -o Sample1.sdf
```

Preparing SD files PubChem deposition:

Consider a SD file with these fields: Mol_ID, Name, Synonyms and Systematic_Name. And Mol_ID data field uniquely identifies your compound.

To prepare a new SD file CmpdDataForPubChem.sdf containing only required PUBCHEM_EXT_DATASOURCE_REGID field, type:

```
% ModifySDFilesDataFields.pl --m datafields
  --datafieldsmap
  "PUBCHEM_EXT_DATASOURCE_REGID,Mol_ID"
  -r CmpdDataForPubChem -o Sample1.sdf
```

To prepare a new SD file CmpdDataForPubChem.sdf containing only required PUBCHEM_EXT_DATASOURCE_REGID field and replace molname line with Mol_ID, type:

```
% ModifySDFilesDataFields.pl --molnamemode datafield
  --molnamereplace always --molname Mol_ID --mode both
  --datafieldsmap
   "PUBCHEM_EXT_DATASOURCE_REGID,Mol_ID"
  -r CmpdDataForPubChem -o Sample1.sdf
```

In addition to required PubChem data field, you can also add optional PubChem data fields.

To map your Name, Synonyms and Systematic_Name data fields to optional PUBCHEM_SUBSTANCE_SYNONYM data field along with required ID field, type:

```
% ModifySDFilesDataFields.pl --molnamemode datafield
  --molnamereplace always --molname Mol_ID --mode both
  --datafieldsmap
  "PUBCHEM_EXT_DATASOURCE_REGID,Mol_ID;
  PUBCHEM_SUBSTANCE_SYNONYM,Name,CompoundName"
  -r CmpdDataForPubChem -o Sample1.sdf
```

To add your <domain.org> as PUBCHEM_EXT_SUBSTANCE_URL and link substance retrieval to your CGI script <http://www.yourdomain.org/GetCmpd.pl,Reg_ID,Mol_ID> via PUBCHEM_EXT_DATASOURCE_REGID field along with optional and required data fields, type:

```
% ModifySDFilesDataFields.pl --molnamemode datafield
  --molnamereplace always --molname Mol_ID --mode both
  --datafieldsmap
  "PUBCHEM_EXT_DATASOURCE_REGID,Mol_ID;
  PUBCHEM_SUBSTANCE_SYNONYM,Name,CompoundName"
  --datafieldscommon
  "PUBCHEM_EXT_SUBSTANCE_URL,domain.org"
  --datafieldURL "PUBCHEM_EXT_DATASOURCE_URL,
  http://www.yourdomain.org/GetCmpd.pl,Reg_ID,Mol_ID"
  -r CmpdDataForPubChem -o Sample1.sdf
```

And to add a publication date and request a release data using PUBCHEM_PUBLICATION_DATE and PUBCHEM_DEPOSITOR_RECORD_DATE data fields along with all the data fields in earlier examples, type: optional fields, type:

```
% ModifySDFilesDataFields.pl --molnamemode datafield
  --molnamereplace always --molname Mol_ID --mode both
  --datafieldsmap
  "PUBCHEM_EXT_DATASOURCE_REGID,Mol_ID;
  PUBCHEM_SUBSTANCE_SYNONYM,Name,CompoundName"
  --datafieldURL "PUBCHEM_EXT_DATASOURCE_URL,
  http://www.yourdomain.org/GetCmpd.pl,Reg_ID,Mol_ID"
  --datafieldscommon
  "PUBCHEM_EXT_SUBSTANCE_URL,domain.org,
  PUBCHEM_PUBLICATION_DATE,YYY-MM-DD,
  PUBCHEM_DEPOSITOR_RECORD_DATE,YYYY-MM-DD"
  -r CmpdDataForPubChem -o Sample1.sdf
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

InfoSDFiles.pl, JoinSDFiles.pl, MergeTextFilesWithSD.pl, SplitSDFiles.pl, SDFilesToHTML.pl

COPYRIGHT