Using the ARFF Output Kettle Plugin

Last Modified on November 26, 2007

Contents

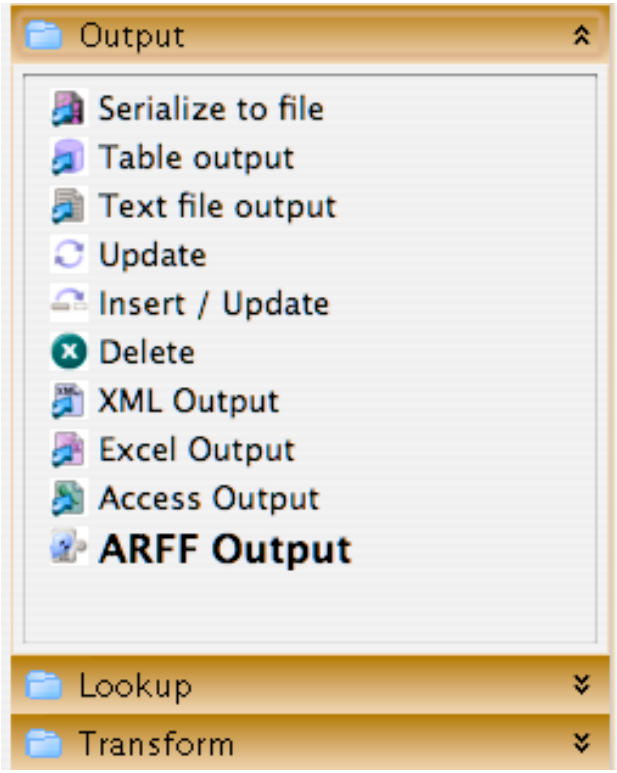
Introduction

The ARFF output plugin is a tool that allows you to output data from Kettle to a file in WEKA's Attribute Relation File Format (ARFF). ARFF format is essentially the same as comma separated values (CSV) format, except with the addition of meta data on the attributes (fields) in the form of a header. Here is an example ARFF file:

```
@relation labor
@attribute 'duration' real
@attribute 'wage-increase-first-year' real
@attribute 'wage-increase-second-year' real
@attribute 'wage-increase-third-year' real
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' real
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' real
@attribute 'shift-differential' real
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' real
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disability-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
@data
1,5,?,?,?,40,?,?,2,?,11,'average',?,?,'yes',?,'good'
2,4.5,5.8,?,?,35,'ret_allw',?,?,'yes',11,'below_average',?,'full',?,'full','goo
d'
?,?,?,?,?,38,'empl_contr',?,5,?,11,'generous','yes','half','yes','half','good'
3,3.7,4,5,'tc',?,?,?,?,'yes',?,?,?,?,'yes',?,'good'
3,4.5,4.5,5,?,40,?,?,?,?,12,'average',?,'half','yes','half','good'
2,2,2.5,?,?,35,?,?,6,'yes',12,'average',?,?,?,?,'good'
. . .
```

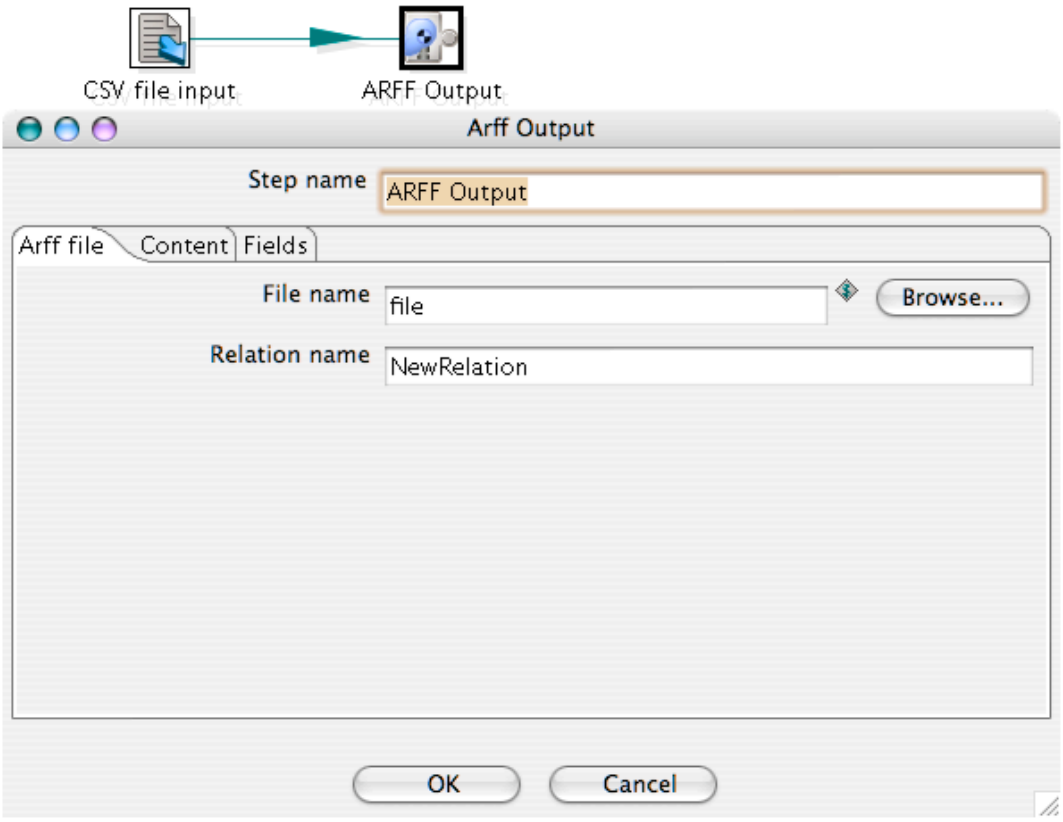
Getting Started

In order to use the ARFF output plugin, it must first be installed correctly in Kettle—unpack the plugin archive and copy all files in the ArffOutputDeploy directory to a sub-directory in the plugins/steps directory of your Kettle installation. You will also need a current version of WEKA available as a library to Kettle. Download WEKA 3.5.6 (or later) from SourceForge, unpack the archive, and copy the "weka.jar" file to the same sub-directory in plugins/steps as before. Now start Spoon. Confirm that the plugin has been installed and correctly recognized by Kettle by expanding the "Output" list under "Core Objects" in Spoon. You should see ARFF output listed in bold near the bottom of the list.



Configuring the ARFF Output Step

Assuming that you have the ARFF output step connected to a previous step, double click on its icon to open the configuration dialog.



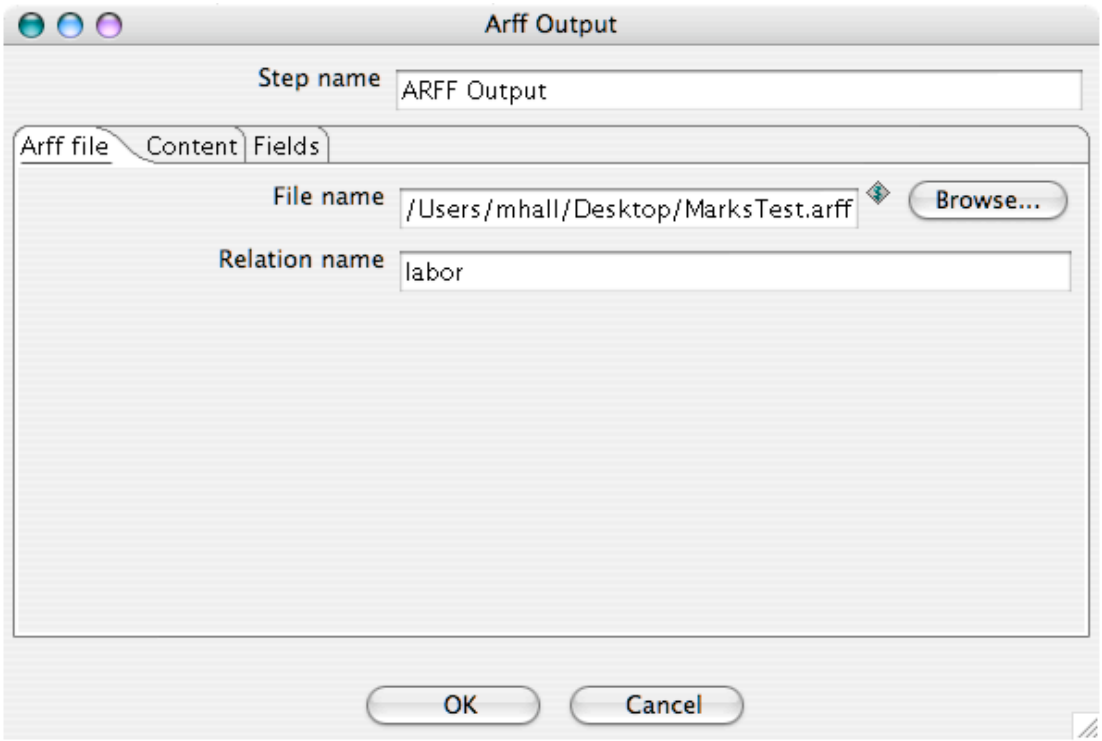
The dialog shows a text field that can be used to name the step and three tabs that can be used to configure the step and review data type mappings.

The steps to complete the configuration are:

1. Choose a file name for the ARFF file
2. Choose a relation name (something short that describes the relationship in the data)
3. Choose a file format to use (if the default is not acceptable)
4. Choose a character encoding to use (if the default is not acceptable)
5. Review the mapping between Kettle field types and ARFF attribute types and correct any problems

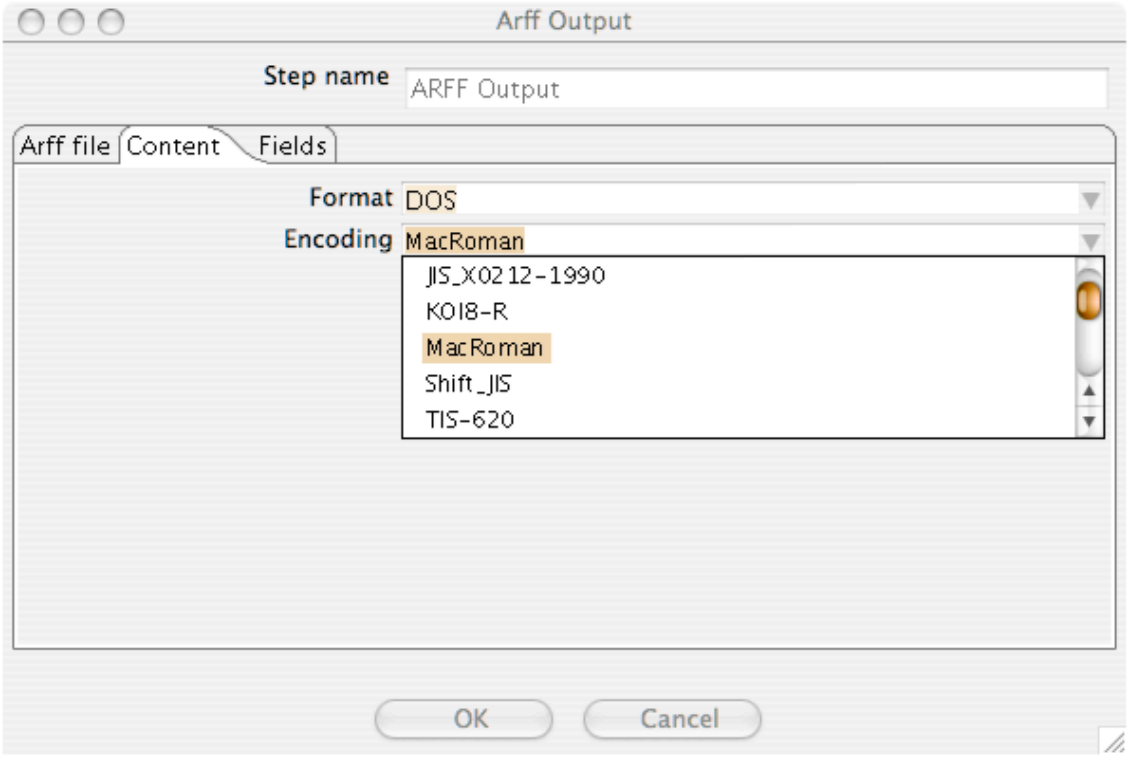
Choosing a Filename and Relation Name

Type a filename into the provided text area on the "File" tab, or use the "Browse" button to navigate through your file system. Enter a relation name for the new ARFF file in the "Relation name" text field.



Choosing a File Format and Character Encoding

File format (Unix or DOS) and character encoding can be selected on the "Content" tab. The default format is DOS and the default character encoding is platform-dependant.



Review the Mapping between Kettle and ARFF types

The mapping between incoming Kettle field types and the ARFF attribute types that will be written to the file is shown in the "Fields" tab. Any obvious problems here can be rectified by modifying the meta data for the Kettle types in a previous step. The ARFF output step maps Kettle "String" type to ARFF "nominal", "Number" and "Integer" to "numeric", and "Date" to "date". All formatting and precision encoded in the Kettle meta data is respected by the ARFF output step.