

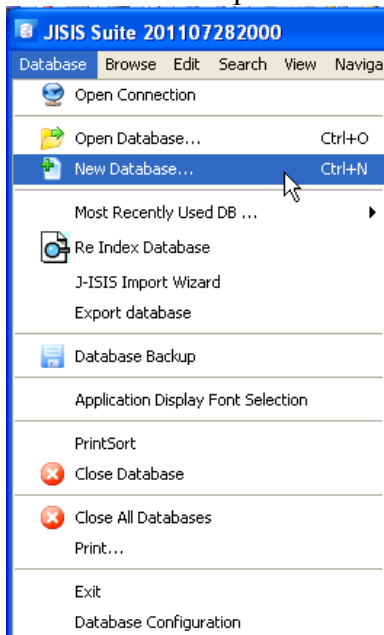
J-ISIS improvements in October 2011 release

The most significant improvements are the following:

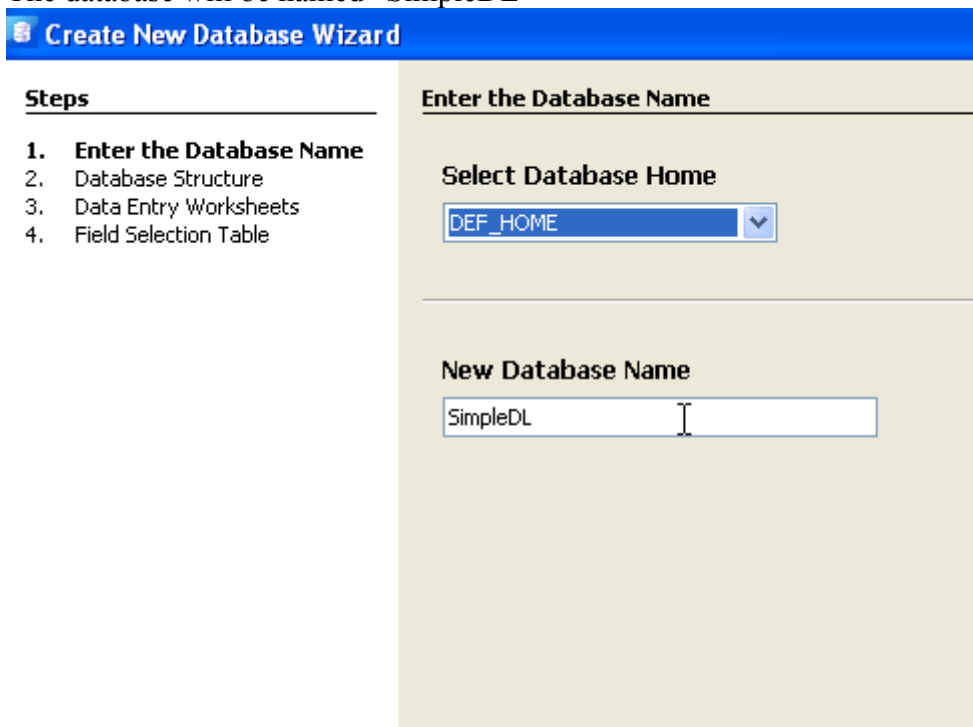
- New Data Entry module for Digital Library management, i.e. you can create a repeatable field with indexing technique 4 and load full documents whatever their format (pdf, doc, docx, rtf, odt, ppt, xls, etc. The document is converted in text and stored in the field first occurrence; an url to the original document is stored in the 2nd occurrence.
- Implementation of stopwords on a file named stopwords.txt for indexing technique 4 (a Spanish version of stopwords.txt would be needed). The file must be stored in the /ifdt folder of the database.
- You can now include HTML references to files that will be loaded in the data viewer and search result panels.
- Latest libraries of Berkeley DB JE (je-4.1.10), Groovy (1.8.3) and Lucene (3.4.0).
- Many bug fixes and speed improvements.

1. New Data Entry module for Digital Library management

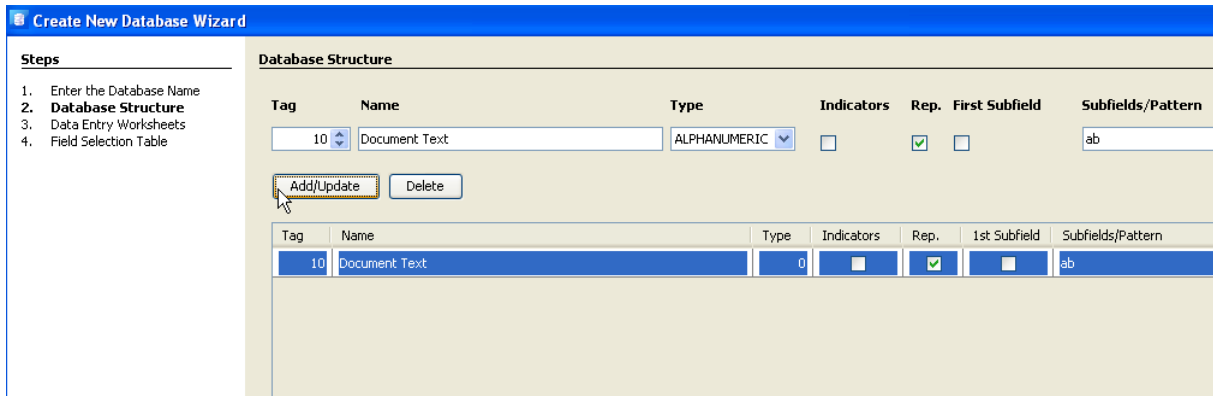
Let's create a simple database by selecting "New Database" in the "Database" menu bar.



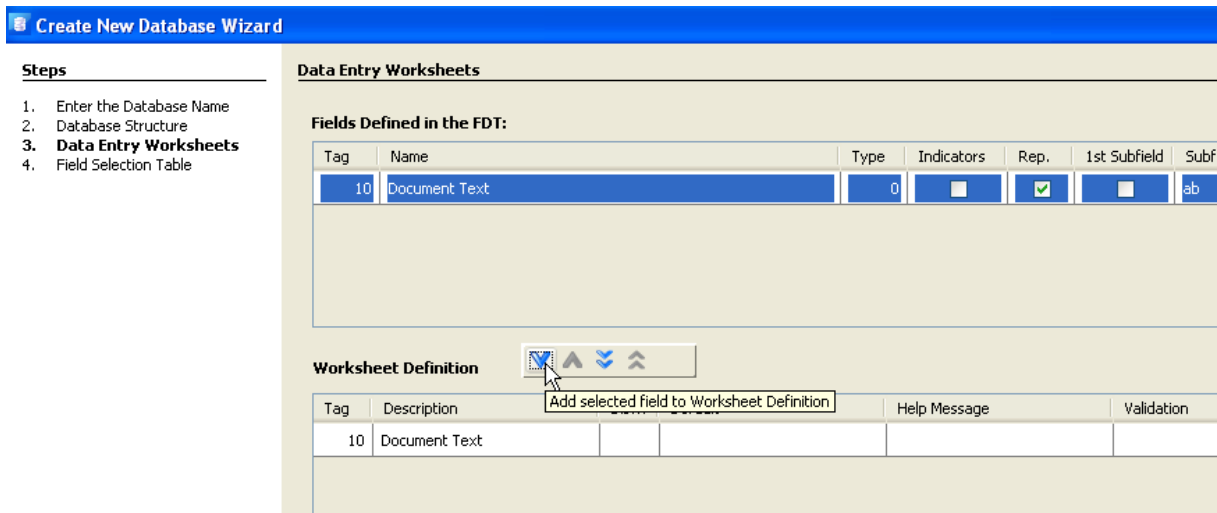
The database will be named "SimpleDL"



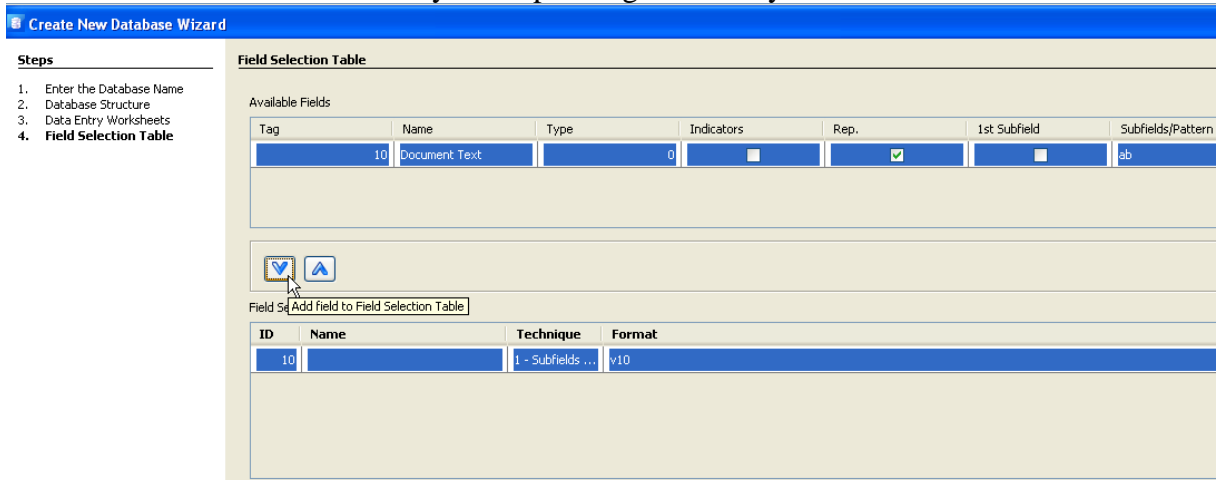
We create a single repetitive field called "Document Text" with tag "10"



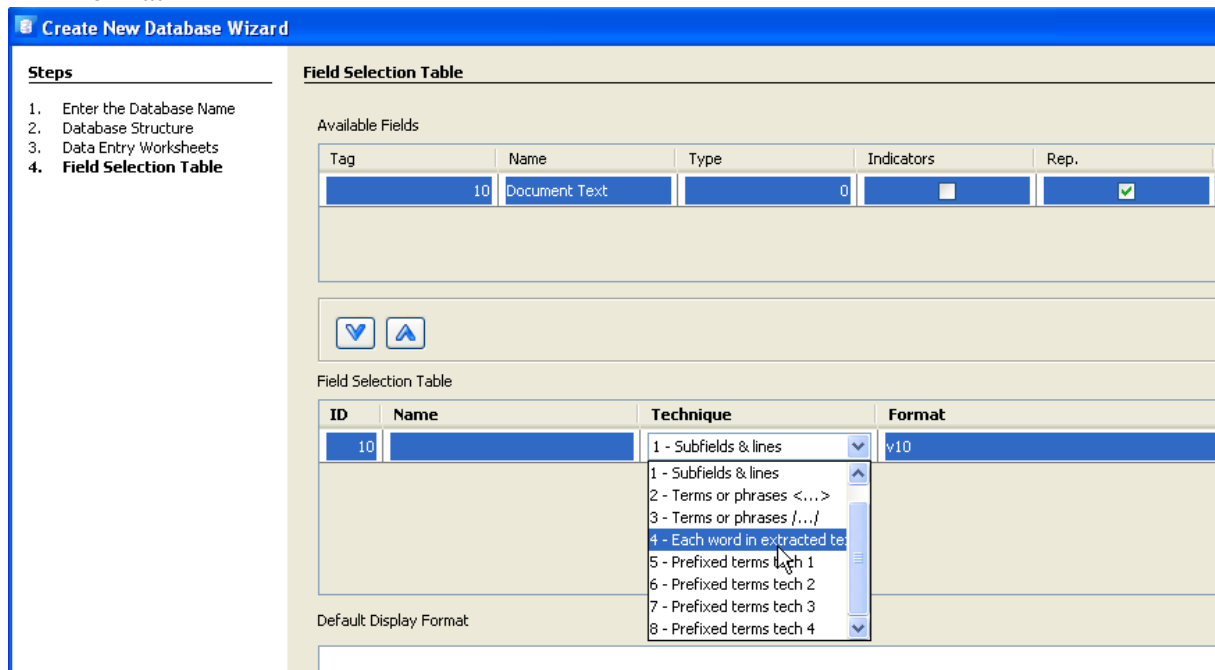
A default worksheet



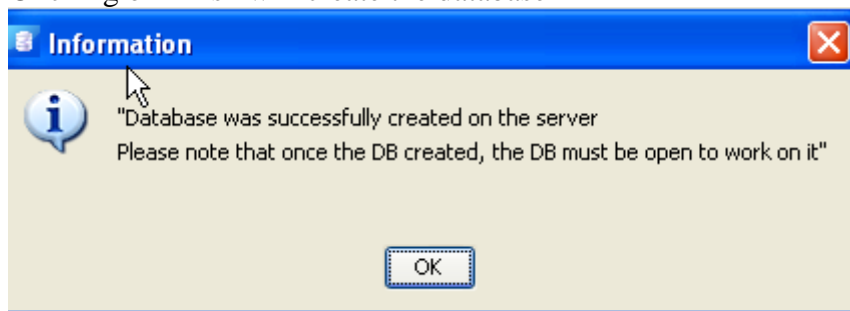
A Field Select Table with an entry corresponding to the only field



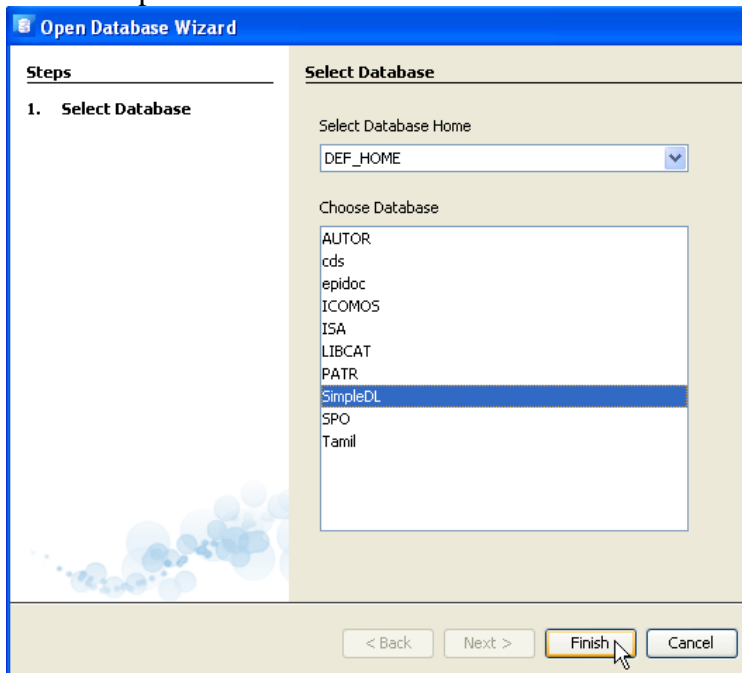
We change the indexing method to 4 which means to index each word in extracted text by the PFT Format



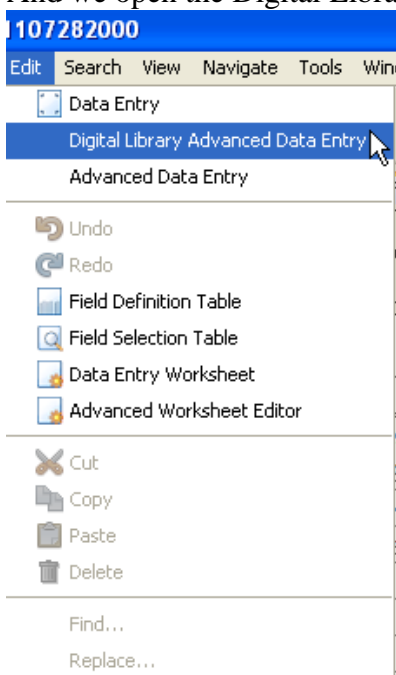
Clicking on finish will create the database



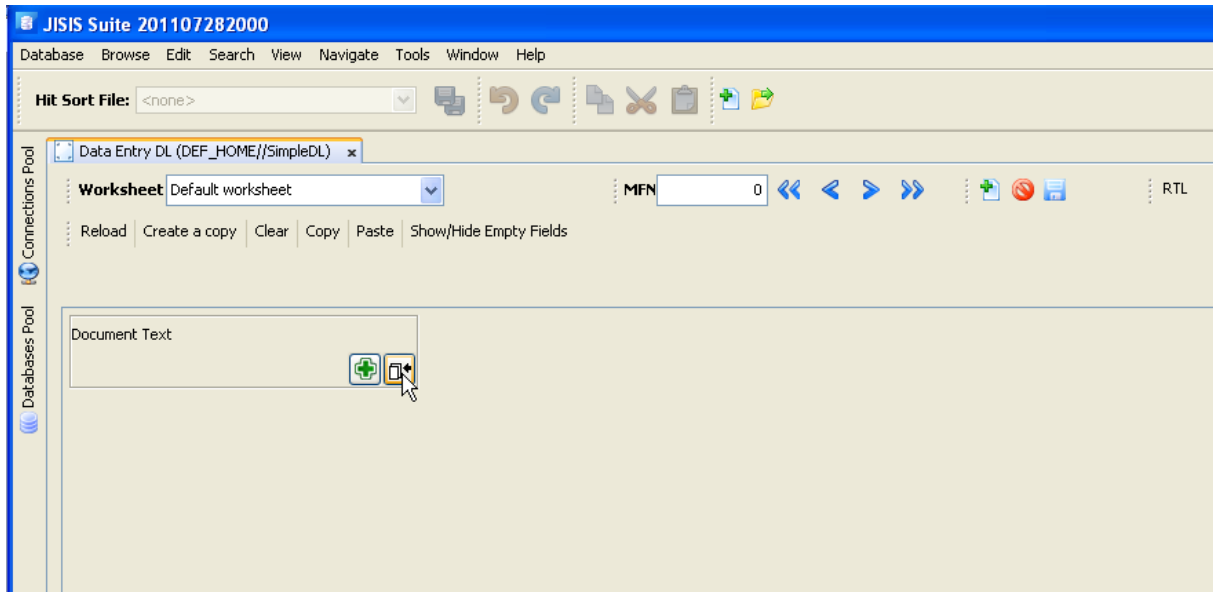
Next we open the new created database



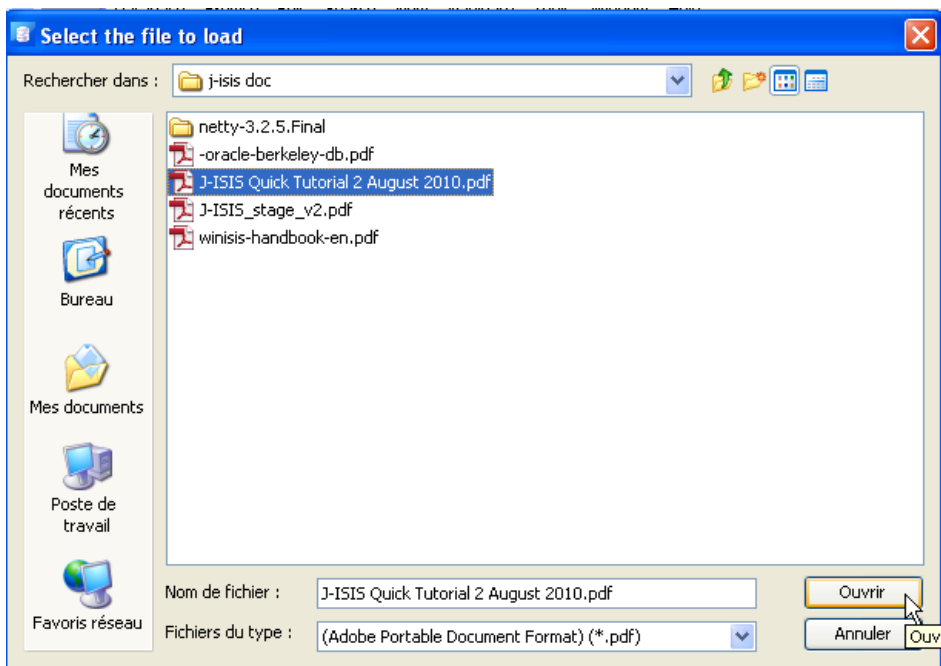
And we open the Digital Library Advanced Data Entry from



The following screen will be displayed. Please note the new button



Clicking on this new button will open a File chooser dialog

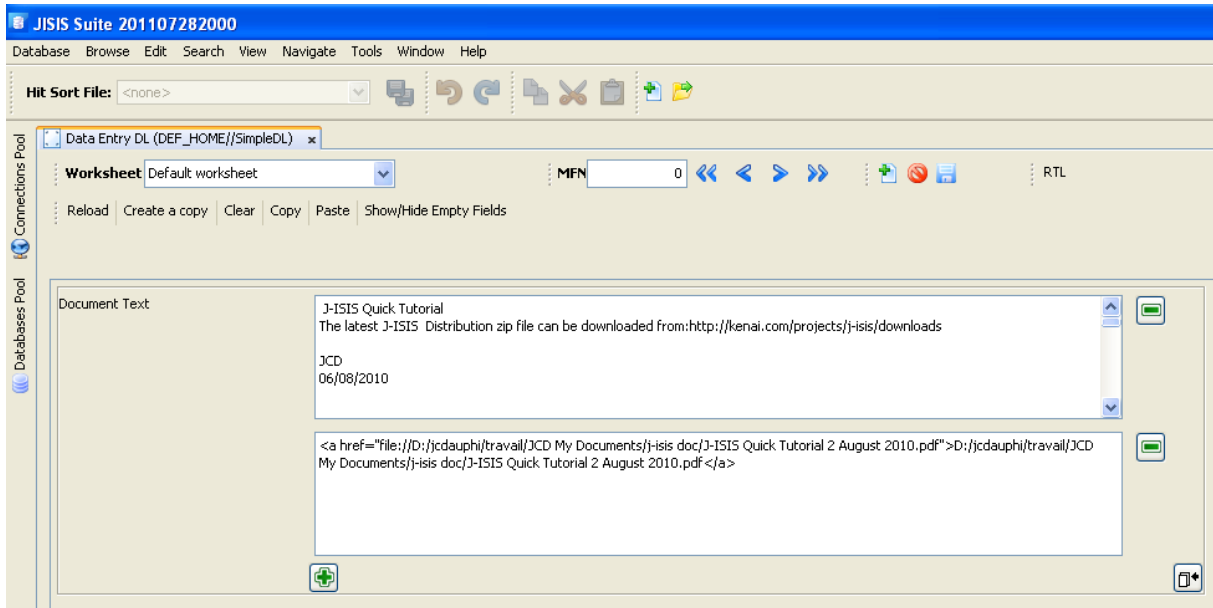


Select the file you wish to load. The following document formats are supported:

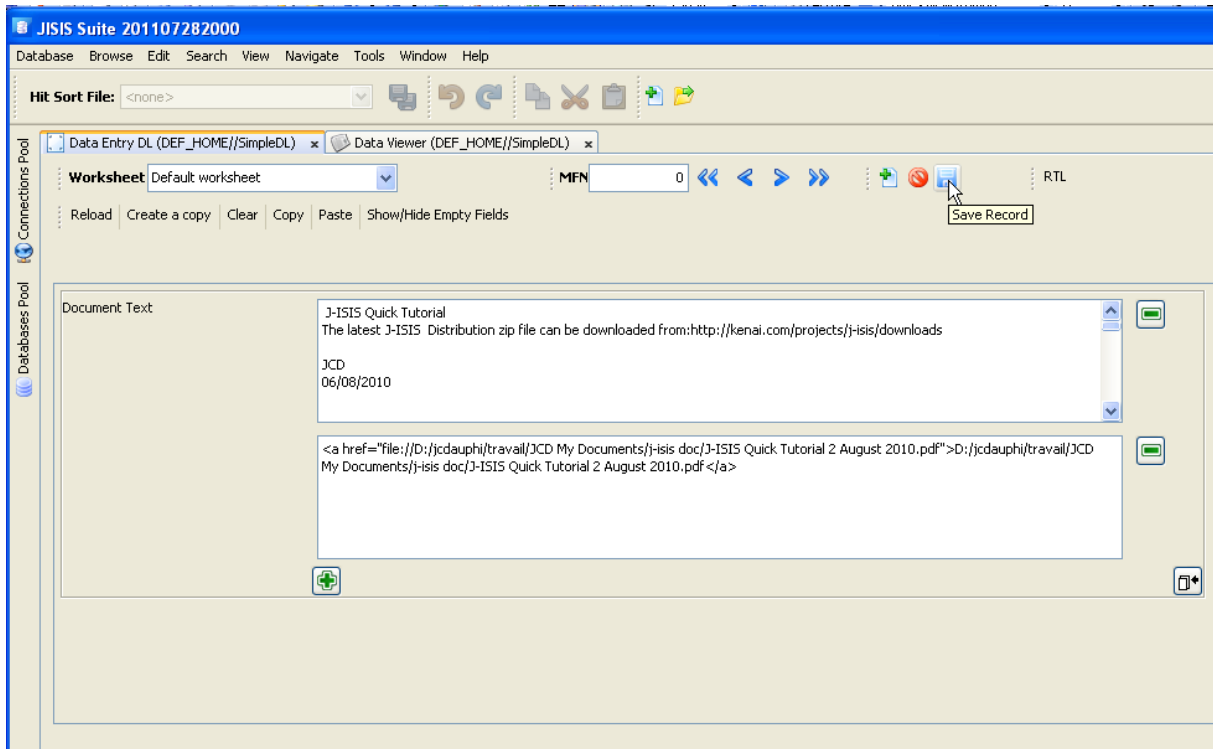
- ("doc", "(MS Word Document")),
- ("pdf", "(Adobe Portable Document Format")),
- ("docx", "(MS docxfiles")),
- ("xls", "(MS Excel Document")),
- ("ppt", "(MS PowerPoint Document")),
- ("rtf", "(Rich Text Format")),
- ("html", "(HTML Format")),

("xhtml", "(XHTML Format)",
("odf", "(OpenDocument)"), "txt", "(Plain Text)").

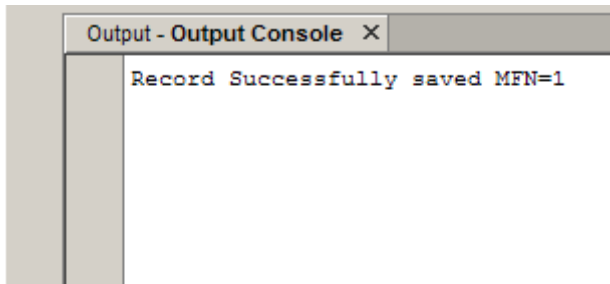
The document will be converted in plain text and displayed in the occurrence where the button was clicked. The next occurrence will contain an url to the original file.



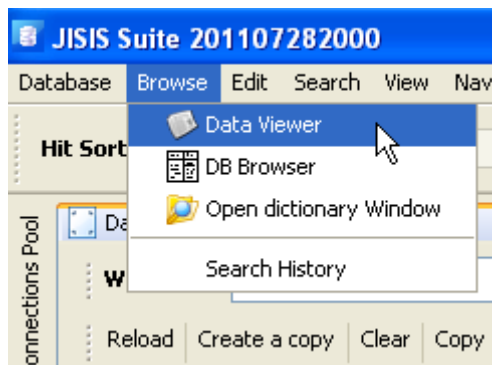
Save the record by clicking on the diskette button.



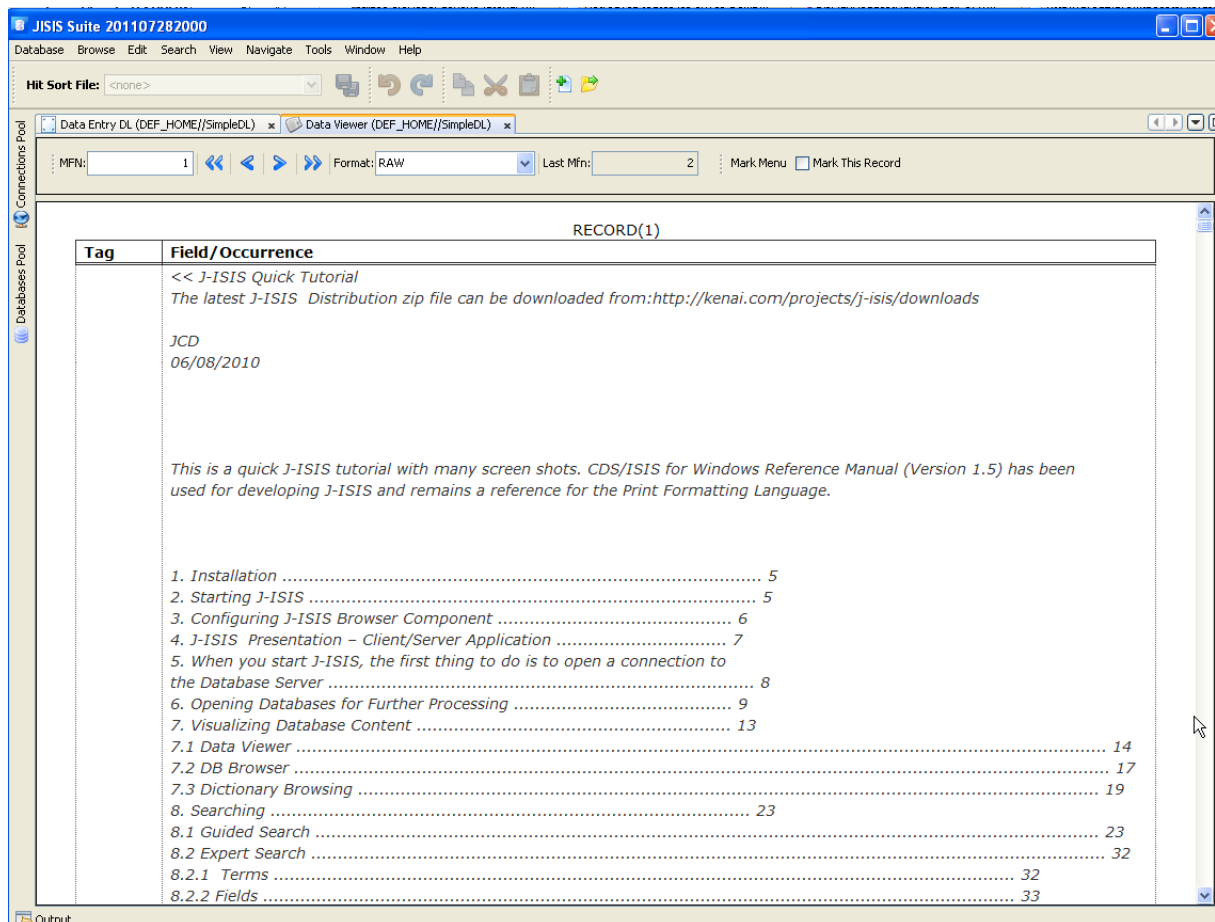
Once record saved and indexed, you will see something like this in the Output window



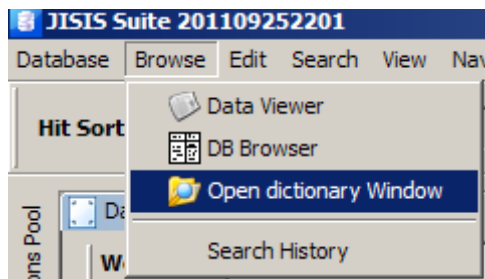
We can now visualize the record through the Data Viewer as follow:



Here is what you should see:



We can also browse the content of the index:



The screenshot shows the Index Content window. The index name is 2011\home_example_db\simpleDL\indexes\master. The number of fields is 2, the number of records is 1, and the number of terms is 2231. The last modified date is Fri Oct 21 18:13:22 CEST 2011. The search options are set to <All Searchable Fields> and the query is empty. The table below shows the indexed terms and their frequencies.

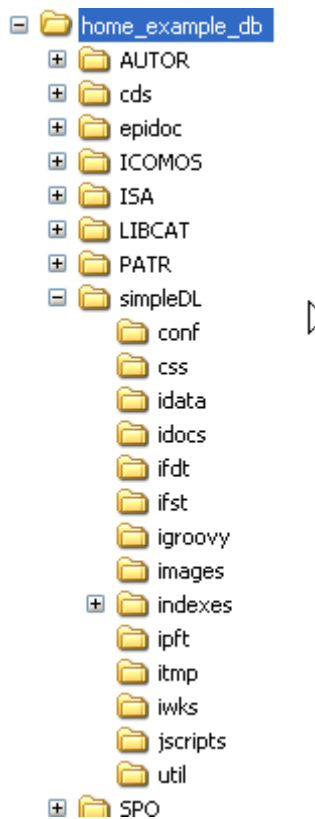
Item	Field	Term	Freq
2071	10	tells	1
2072	10	template	1
2073	10	templates	1
2074	10	term	1
2075	10	terms	1
2076	10	test	1
2077	10	tested	1
2078	10	tester	1
2079	10	testing	1
2080	10	testjisis15	1
2081	10	tests	1
2082	10	text	1
2083	10	than	1
2084	10	them	1
2085	10	therefore	1
2086	10	thereof	1
2087	10	thing	1
2088	10	think	1
2089	10	third	1
2090	10	those	1
2091	10	three	1
2092	10	through	1

We can see that on significant terms are indexed such as “than”, “them”, “the”

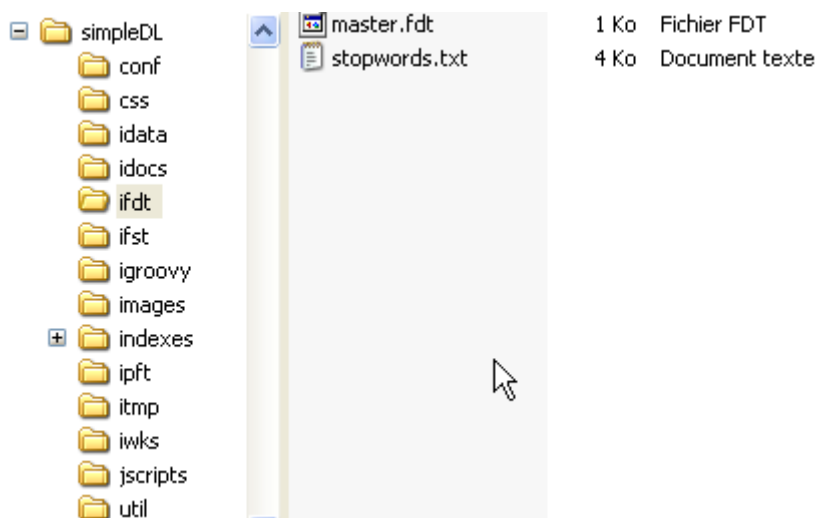
2. Indexing by eliminating stopwords

The stopwords file needs to be set up outside J-ISIS using a text editor or word processor. It must have **stopwords.txt** as file name and it must reside in the ifdt folder as the FDT file for the database.

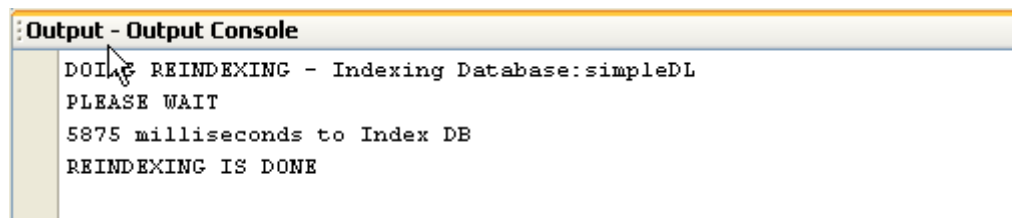
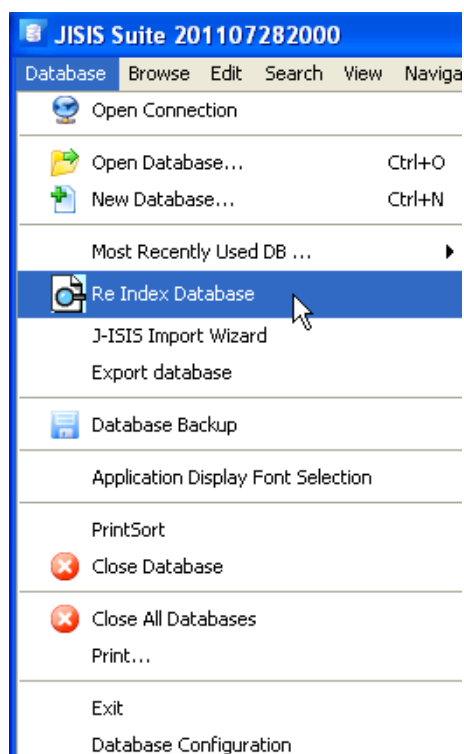
Database folder organization



Placing the “stopwords.txt” file in the /ifdt folder



Then, we re-index the database



Looking again at the index, we can see that the stopwords have been eliminated

The screenshot shows the JISIS Suite 201107282000 interface. The main window displays the 'Index Content' for the index 'org.apache.lucene.store.SimpleFSDirectory@D:\jcd'. The index statistics are as follows:

- Index name: org.apache.lucene.store.SimpleFSDirectory@D:\jcd
- Number of fields: 2
- Number of Records: 1
- Number of terms: 1914
- Last modified: Fri Oct 21 18:21:38 CEST 2011

A 'Quick Search' section is present with a dropdown menu set to '<All Searchable Fields>' and an empty 'Query' input field.

The 'Index Content' table is as follows:

iTerm	Field	Term	Freq
1798	10	tags	1
1799	10	task	1
1800	10	te	1
1801	10	technique	1
1802	10	techniques	1
1803	10	tells	1
1804	10	template	1
1805	10	templates	1
1806	10	term	1
1807	10	terms	1
1808	10	test	1
1809	10	tested	1
1810	10	tester	1
1811	10	testing	1
1812	10	testjisis15	1
1813	10	tests	1
1814	10	text	1
1815	10	thereof	1
1816	10	thing	1
1817	10	throws	1
1818	10	tides	1
1819	10	tilde	1

In the “Data Viewer”, we can see the 2nd occurrence of the Document Text field that display a url to the original document.

The screenshot shows the JISIS Suite 201107282000 interface. The title bar reads "JISIS Suite 201107282000". The menu bar includes "Database", "Browse", "Edit", "Search", "View", "Navigate", "Tools", "Window", and "Help". Below the menu bar is a toolbar with icons for file operations. The "Hit Sort File:" dropdown is set to "<none>". The window title bar shows three tabs: "Data Entry DL (DEF_HOME//simpleDL)", "Data Viewer (DEF_HOME//simpleDL)", and "Dictionary (DEF_HOME//simpleDL)". The "Data Viewer" tab is active, showing a record with MFN: 1 and Last Mfn: 2. The format is set to "RAW". The main display area shows the following text:

```
133
We create an instance of the pdfCatalogue class and we call the process method:

def catalogue = new pdfCatalogue()
catalogue.process()

Click on the "Execute Groovy Script" Toolbar button to execute pdfCatalogue script.

During execution, you should see the following dialog:

And when the dialog disappears, you should see:

134

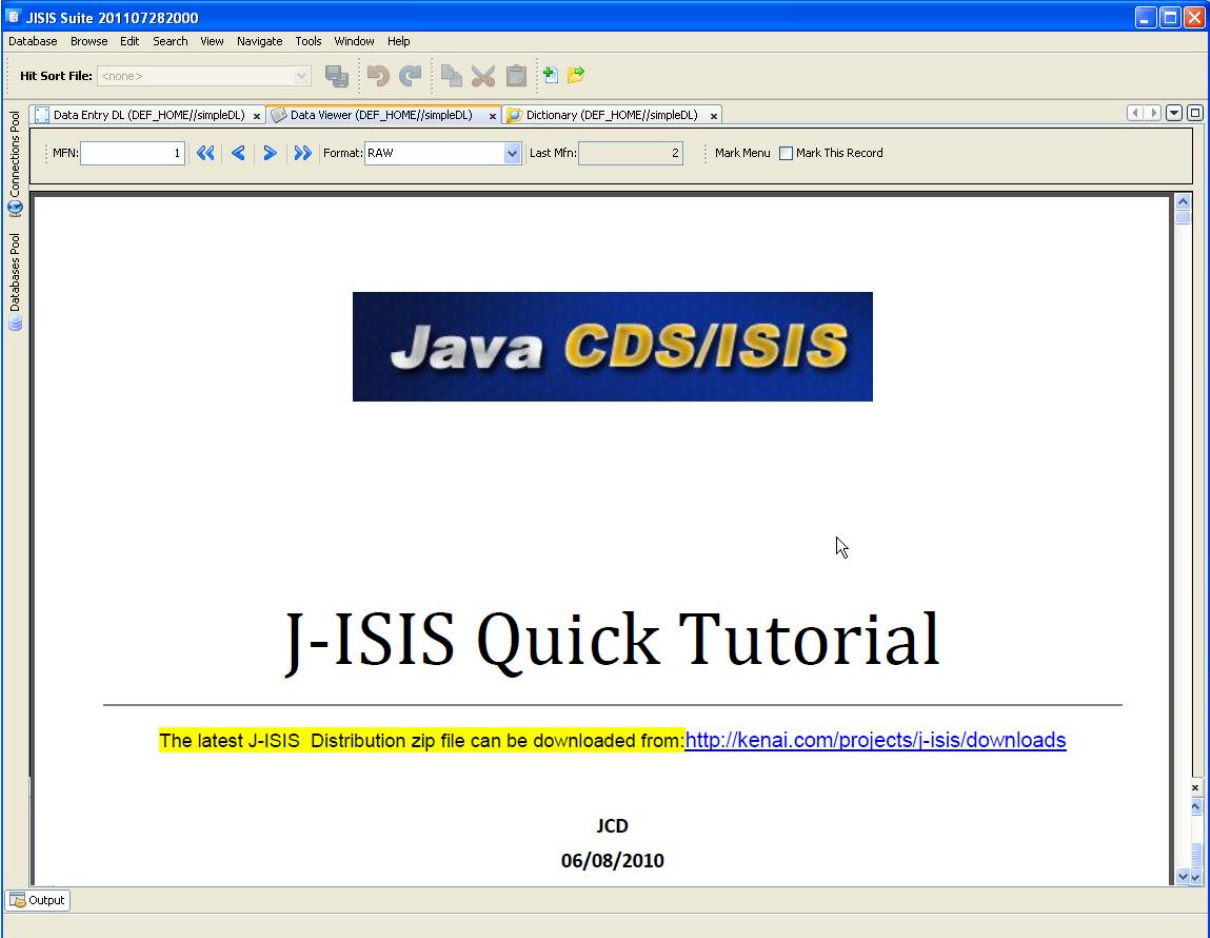
If you don't provide a full path, the output file "asfaex.pdf" will be stored in the j-isis root folder it it should
look like:

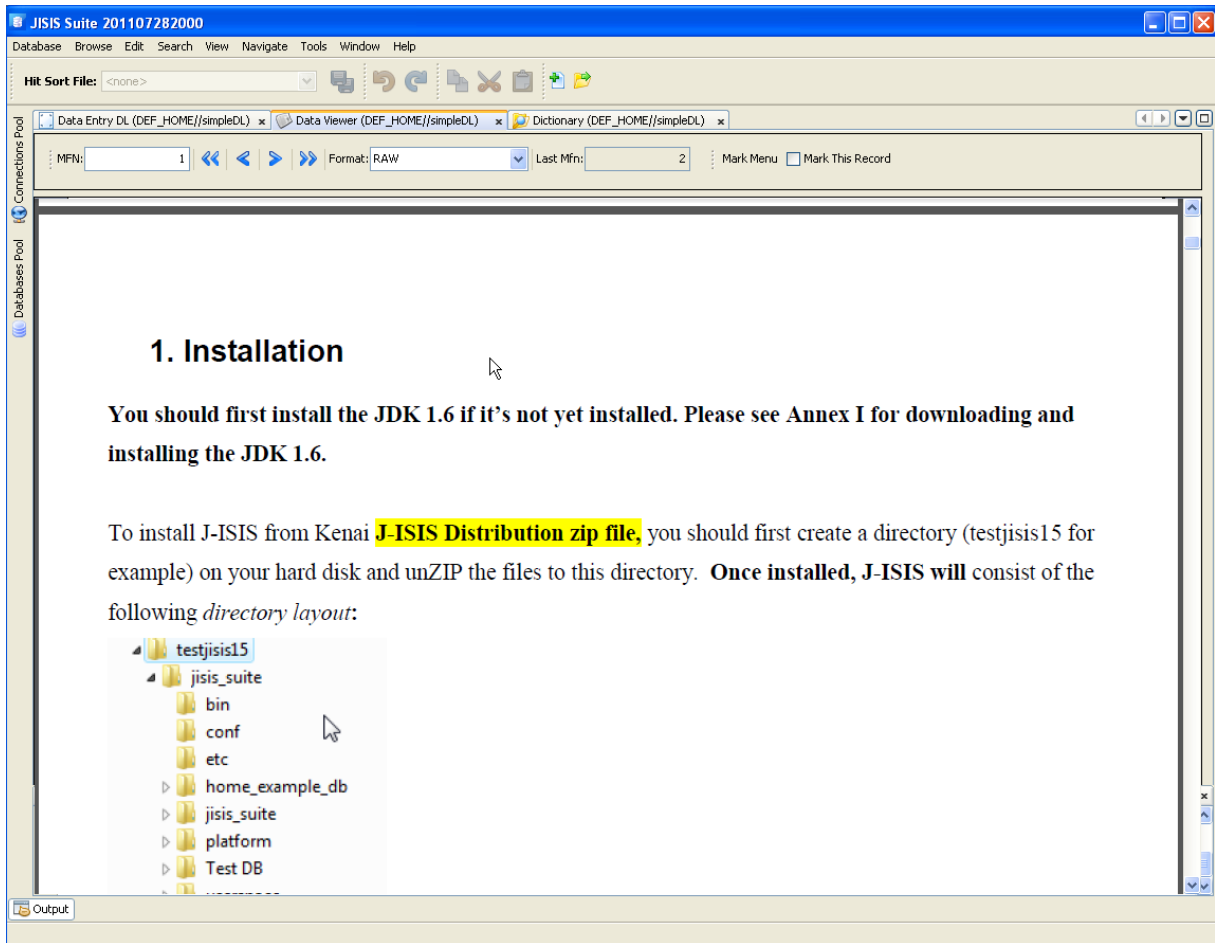
135

136

>>
10: <<D:/jcdaphi/travail/JCD My Documents/j-isis doc/J-ISIS Quick Tutorial 2 August 2010.pdf>>
```

Clicking on the url will load the document in the Window panel.





3. STOPWORDS LIST

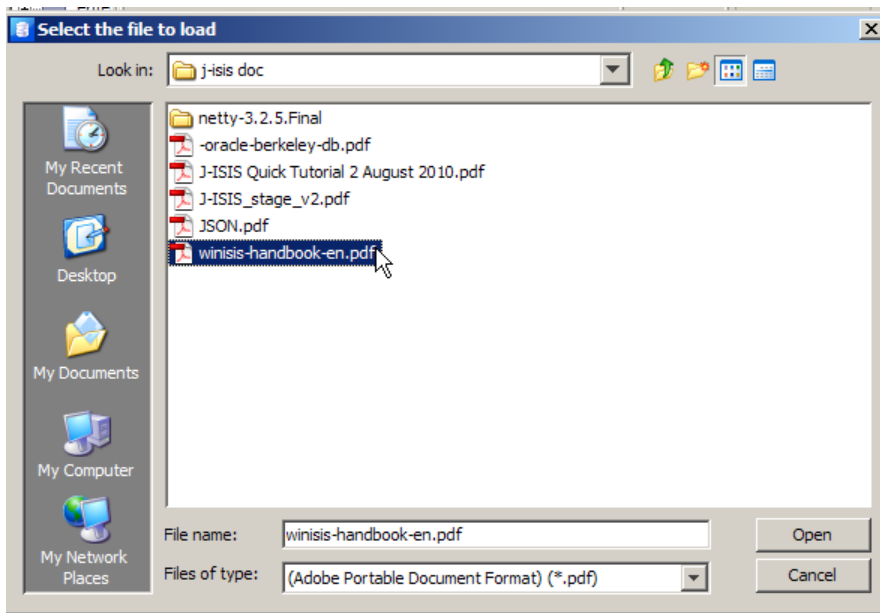
When indexing a field by separate words (indexing technique 4) it may be necessary to prevent common, non-informative words such as 'an' or 'the' from being indexed. This can be achieved by setting up a *stopword list* for the database. Words on the stopword list will not be indexed using indexing techniques 4 (though they may still appear as part of phrases produced with other indexing techniques). Note that there can only be one stopword list for a given database, not different lists for different fields.

The stopword file needs to be set up outside J-ISIS using a text editor or word processor. It must have **stopwords.txt** as file name and it must reside in the ifdt folder as the FDT file for the database.

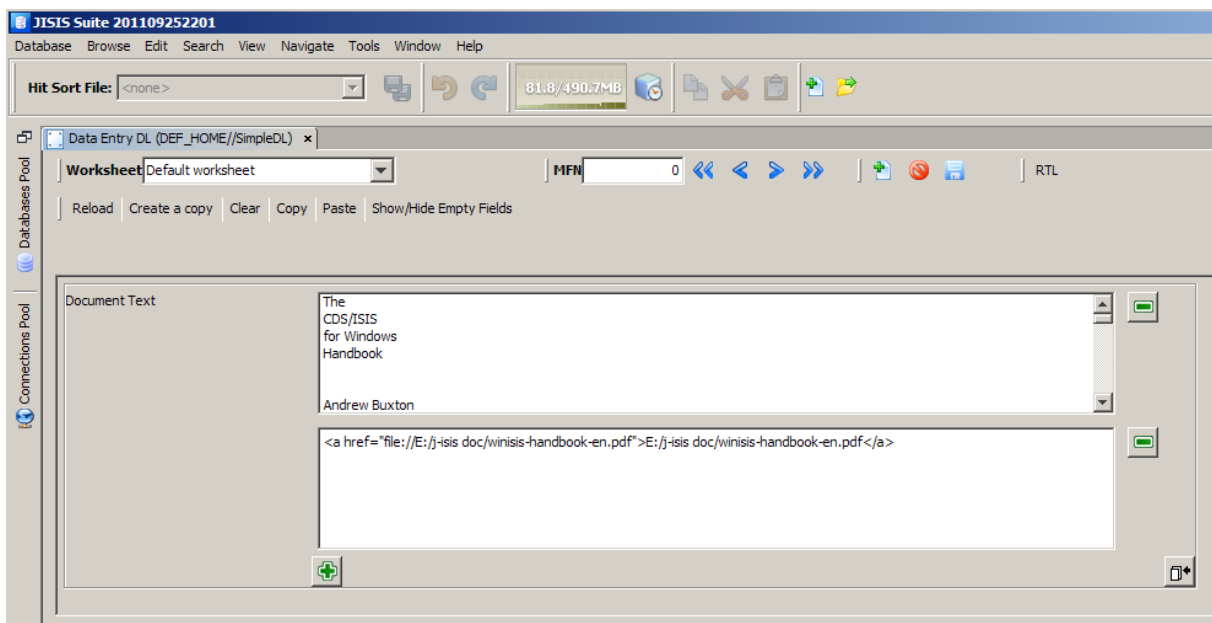
The file must contain one stopword on each line with no preceding spaces, the recognition of stopwords is case insensitive and the words. An example is shown below.

a
able
about
above
according
accordingly
across
actually
after
afterwards
again
against
all

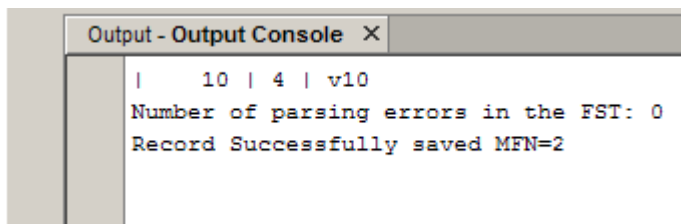
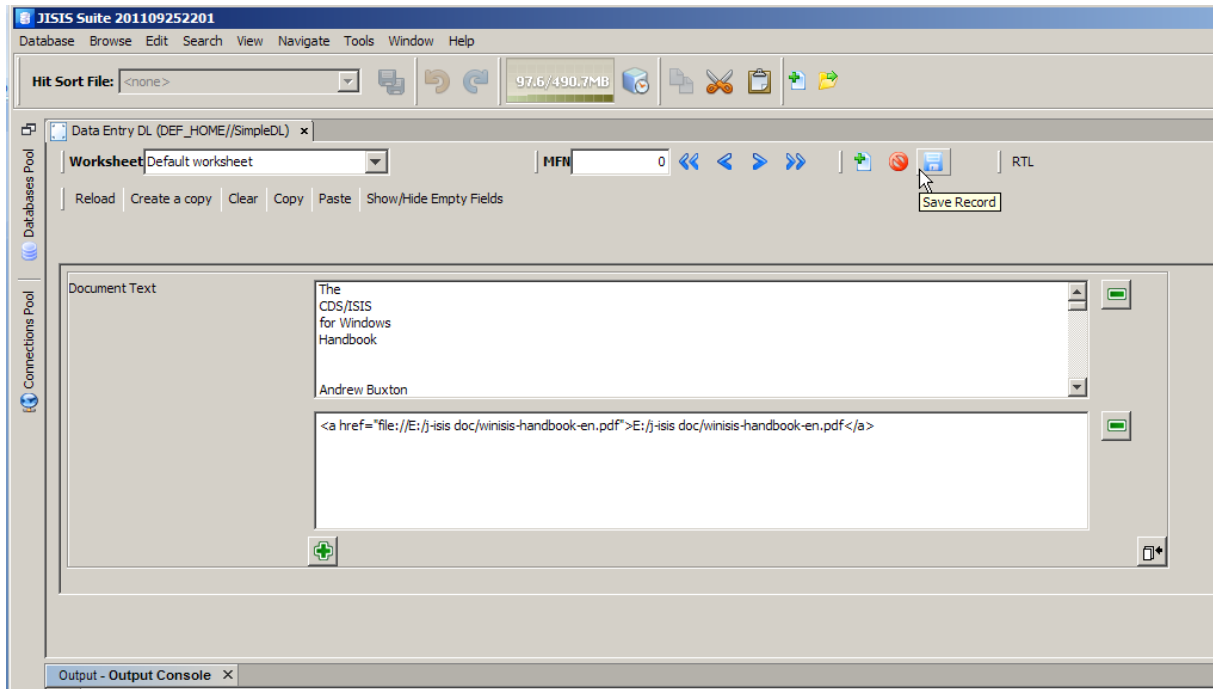
4. Examples of documents that can be loaded and indexed



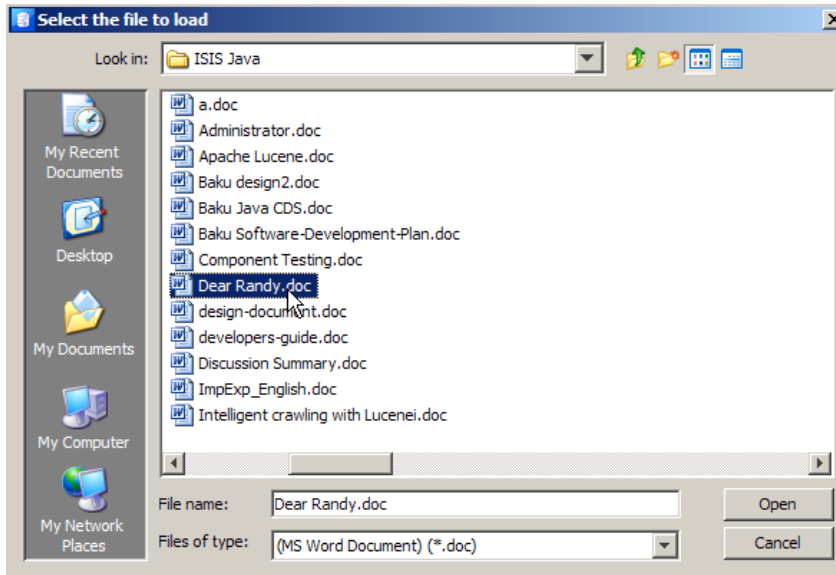
The document is loaded

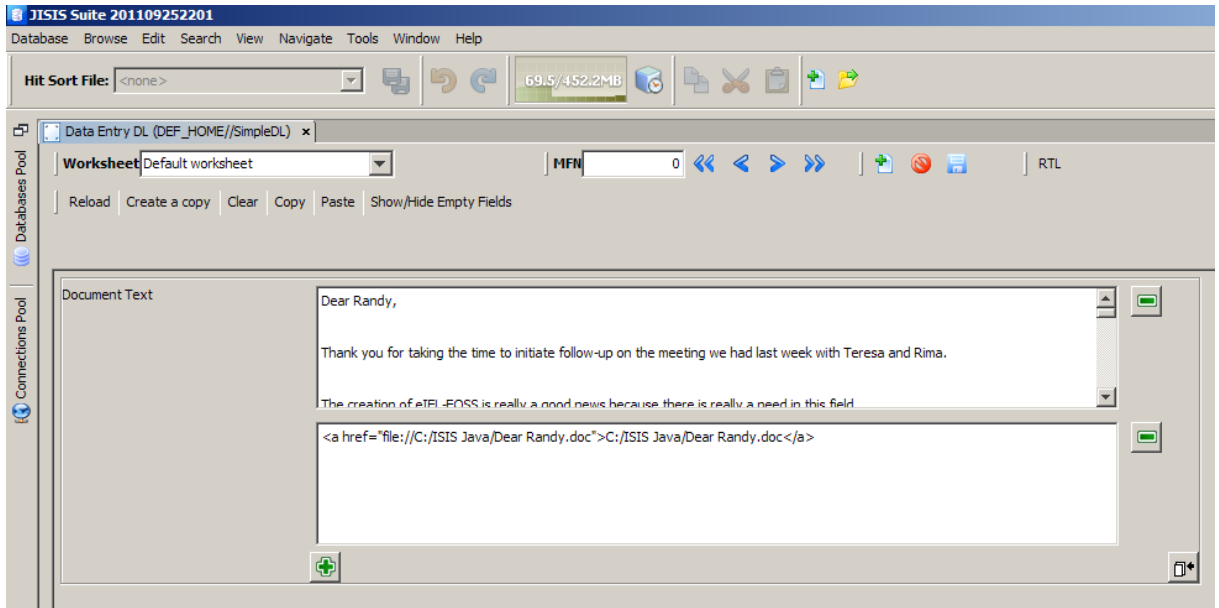


And saved



Let's load a word document





And save it

