INSTALLING AND RUNNING THE VIZLINC INGESTER

INTRODUCTION

The VizLinc ingester is a demonstration tool that processes a text documents, performing named-entity recognition, geocoding, and social network construction. It produces the databases needed to browse and search the documents and data using the VizLinc data visualization tool.

The named-entity recognizer model supplied with the ingester is trained to find English entities that are Locations, Organizations, and Persons. The Ingester will geocode locations that are countries of the world and, for demonstration purposes, named places in the country in Colombia. The Ingester also generates a social network graph of all people who appear in at least two documents. The graph includes an edge between two people if they appear together in at least two documents.

REQUIREMENTS

Like the VizLinc tool itself, the VizLinc ingester has been tested on Windows 7 SP1 (64 bit), using 64-bit Oracle Java 1.7.0_45 and 1.7.0_51.

INSTALLATION

The Ingester is supplied in the file named **vizlinc-ingester.zip**. Unzip this file anywhere you wish, and end up with a folder named **vizlinc-ingester**.

INPUTS

The ingester takes as input folder of files that contain textual information. The folder may contain a hierarchy of subfolders, which are also processed. The ingester extracts text from the files, which may be Microsoft Office files (.doc, .docx, .xls, .xlsx, etc.), PDF files (with embedded text), plaintext, XML, and many other formats. For a list of all supported formats, see https://tika.apache.org/1.4/formats.html.

The ingester also asks you for an output folder, and a prefix for the output files and directory (see below).

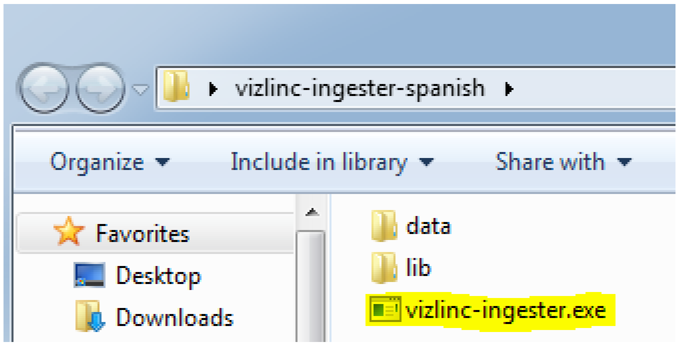
OUTPUTS

The ingester produces as output files and folders, which are used as input to the Vizlinc data visualization tool. All of these file and directory names share a common prefix that you supply. The files and folders are:

prefix.**h2**	database of named entity and document information
prefix.**lucene**	search index for document keyword search
prefix.**graphml**	social network graph

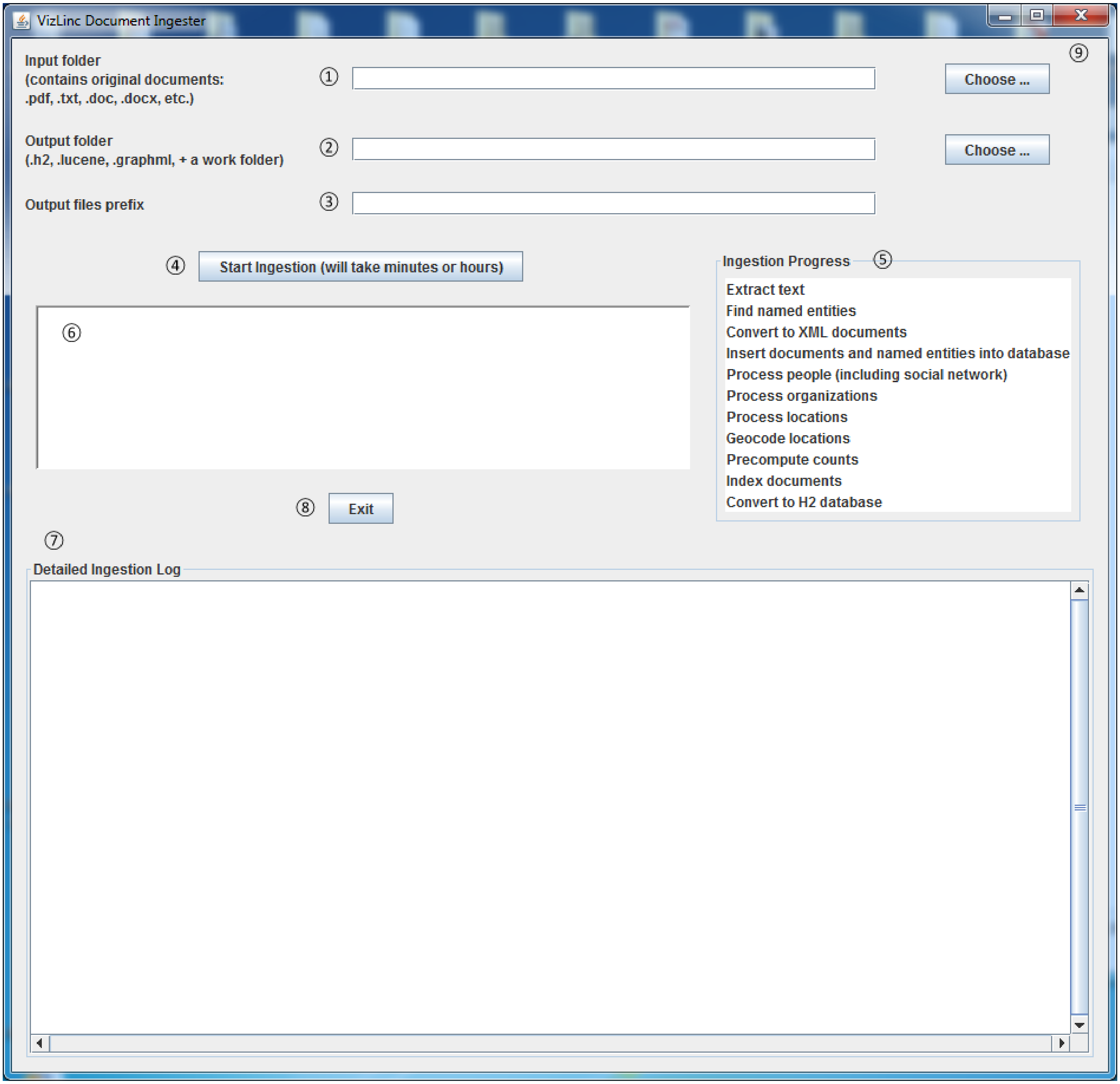
OPERATION

To run the ingester, open the **vizlinc-ingester** folder and double-click on the **vizlinc-ingester.exe** file.



A splash screen will display as the ingester starts up:

Then this window will appear (see next page):



In the **Input folder** field ①, type the full pathname for the folder that contains the text documents you want to process. Or use the **Choose…** button to select the folder.

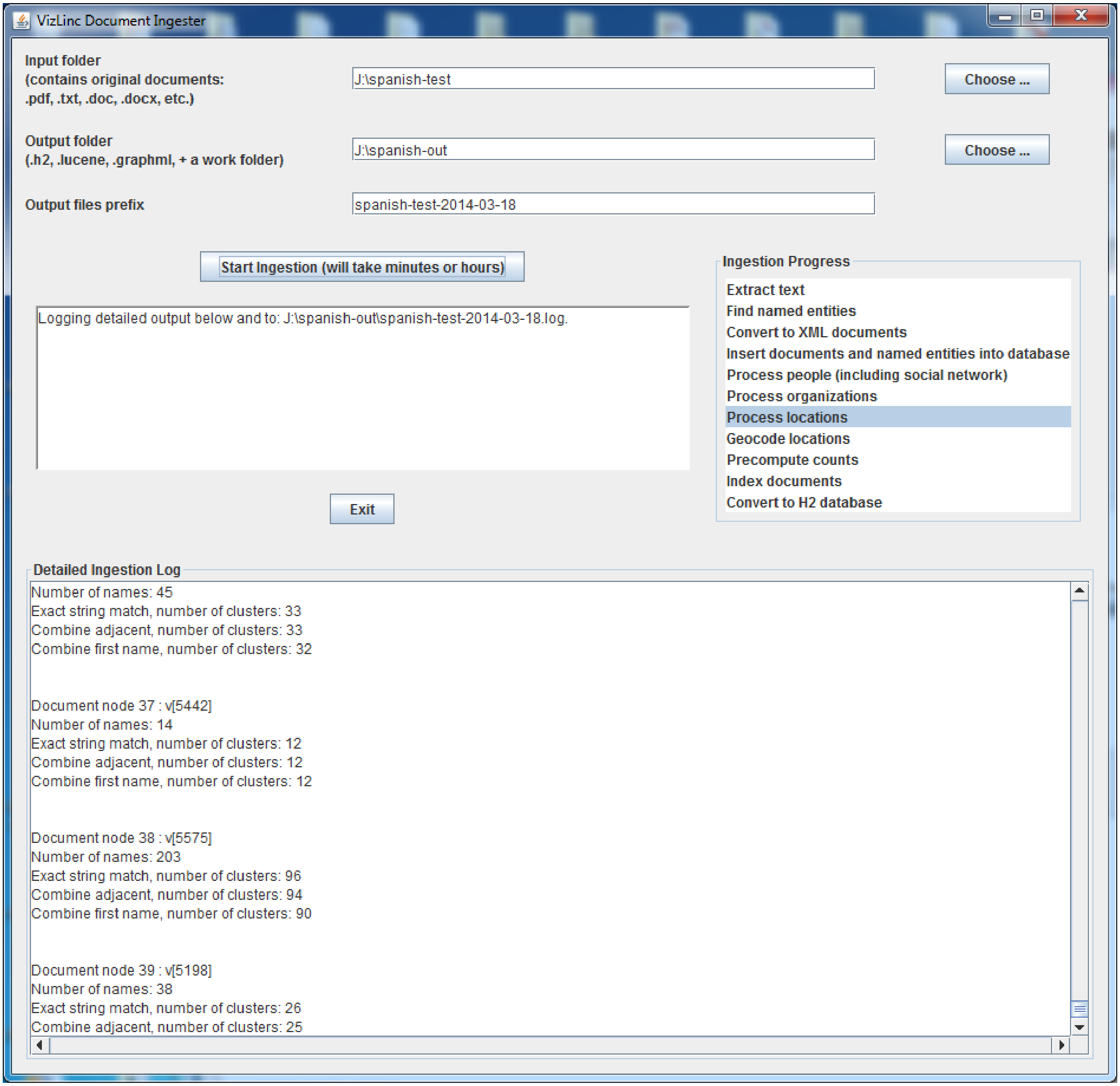
In the **Output folder** field ②, type the full pathname for the folder where you would like the output files and folders to be placed. The folder does not need to exist: it will be created. Or use the **Choose…** button to select the folder

In the **Output files prefix** field ③, type the prefix you want to use to name the files and folders that will be created as described in the **OUTPUTS** section above. You can use any prefix you want as long as it can be part of a filename. You might use a prefix like "mydocs-2014-02-20", for instance.

Finally, press the **Start Ingestion** button ④. The ingester will start processing the input files. As it goes through its various stages, the ingester will highlight the appropriate line in the **Ingestion Progress** box ⑤. It will write detailed progress information to a log file whose location will be given to you in the empty box below the button ⑥. The ingester will also write the same log information to the large area labeled **Detailed Ingestion Log** ⑦.

When the ingestion has finished, press the **Exit** button ⑧ or close the window with the **X** in the upper right corner ⑨ to stop the ingester.

Here's an example of the ingester window while ingestion is in progress:


VizLinc Document Ingester
Choose ...
Choose ...
Start Ingestion (will take minutes or hours)
Exit

Input folder
(contains original documents:
.pdf, .txt, .doc, .docx, etc.) J:\spanish-test

Output folder
(.h2, .lucene, .graphml, + a work folder) J:\spanish-out

Output files prefix spanish-test-2014-03-18

Logging detailed output below and to: J:\spanish-out\spanish-test-2014-03-18.log.

Ingestion Progress

Extract text
Find named entities
Convert to XML documents
Insert documents and named entities into database
Process people (including social network)
Process organizations
Process locations
Geocode locations
Precompute counts
Index documents
Convert to H2 database

Detailed Ingestion Log

```
Number of names: 45
Exact string match, number of clusters: 33
Combine adjacent, number of clusters: 33
Combine first name, number of clusters: 32


Document node 37 : v[5442]
Number of names: 14
Exact string match, number of clusters: 12
Combine adjacent, number of clusters: 12
Combine first name, number of clusters: 12


Document node 38 : v[5575]
Number of names: 203
Exact string match, number of clusters: 96
Combine adjacent, number of clusters: 94
Combine first name, number of clusters: 90


Document node 39 : v[5198]
Number of names: 38
Exact string match, number of clusters: 26
Combine adjacent, number of clusters: 25
```