Instructions for working with the output of the extracted USAID pdf files

Files supplied:

- dec_metadata.json: A collection of the metadata for the text files
- file_halves.json : A collection of arrays of txt. Each array represents one half of a text file, with one line per array entry.

Relationship between collections:

The **lower_id** key in the metadata collection is equivalent to the **file_id** key in the text collection.

Working with the files:

It is recommended that the files be imported into a NoSQL database such as MongoDB, CouchDB, ElasticSearch.

Importing files into MongoDB:

The files can be imported using the mongoimport utility. More information can be found here: http://docs.mongodb.org/manual/reference/program/mongoimport/

The commands would be something like:

```
mongoimport --db dec --collection dec_files --file file_halves.json
mongoimport --db dec --collection dec_metadata --file dec_metadata.json
```

Importing files into CouchDB

The commands would be something like:

```
curl -d @file_halves.json -H "Content-type: application/json" -X POST
http://127.0.0.1:5984/dec/_bulk_docs
curl -d @dec_metadata.json -H "Content-type: application/json" -X POST
http://127.0.0.1:5984/dec/_bulk_docs
```

Importing files into ElasticSearch

The commands would be something like:

```
curl -XPUT localhost:9200/_bulk --data-binary @file_halves.json
curl -XPUT localhost:9200/_bulk --data-binary @file_halves.json
```

Resources for learning more about JSON databases:

- http://en.wikipedia.org/wiki/NoSQL
- http://www.mongodb.com/nosql-explained

Resources for learning more about JSON:

- http://www.json.org/

- http://www.copterlabs.com/blog/json-what-it-is-how-it-works-how-to-use-it/