IRCO Documentation

Release 0.10.2

Jonathan Stoppani

August 23, 2014

Contents

1 Installation **1**
- 1.1 Requirements . . . 1
- 1.2 Installing . . . 1
- 1.3 Upgrading . . . 1

2 Quick Start **3**
- 2.1 Initializing the database . . . 3
- 2.2 Importing some data . . . 3
- 2.3 Browsing the database . . . 3
- 2.4 Generating a graph . . . 3

3 Configuration **5**
- 3.1 Initialization on OS X . . . 5

4 Supported input formats **7**
- 4.1 Compendex . . . 7
- 4.2 Scopus . . . 7
- 4.3 Web of Science . . . 7

5 Generating graphs **9**
- 5.1 Graph types . . . 9
- 5.2 Filtering publications . . . 9
- 5.3 Misc notes . . . 10

6 Command line utilities reference **13**
- 6.1 `irco-convert` . . . 13
- 6.2 `irco-explorer` . . . 13
- 6.3 `irco-geolocate` . . . 13
- 6.4 `irco-graph` . . . 13
- 6.5 `irco-import` . . . 13
- 6.6 `irco-init` . . . 13

7 Downloading search results from WoS **15**
- 7.1 Shell script . . . 15
- 7.2 Bookmarklet . . . 15

8 Backing up the database **17**
- 8.1 SQLite 3 . . . 17

9 News **19**
9.1 Release notes 19

10 Indices and tables **23**

CHAPTER 1

Installation

1.1 Requirements

Todo

1.2 Installing

The complete IRCO suite, including its requirements, can be installed though `pip` by issuing the followin command:

```
pip install irco
```

1.3 Upgrading

It is possible to upgrade the IRCO tool to its latest version by issuing the following command:

```
pip install --upgrade irco
```

CHAPTER 2

Quick Start

2.1 Initializing the database

The first step to start using the IRCO suite is to create an empty database to hold the publication data. This can be done by executing the `irco-init` command as shown below:

```
irco-init sqlite:///irco.db
```

2.2 Importing some data

Todo

2.3 Browsing the database

Todo

2.4 Generating a graph

Todo

CHAPTER 3

Configuration

IRCO uses INI files as its configuration format. Each time an IRCO command is run, configuration is loaded from different places:

- `/etc/irco/irco.ini`
- `~/.irco.ini`
- `irco.ini` in the directory where the command is executed

Values defined in following files overwrite the same directives defined in previous files.

3.1 Initialization on OS X

To create a new user-specific configuration file, issue the following commands:

```
cd ~
touch .irco.ini
open -a TextEdit .irco.ini
```

At this point, the TextEdit application opens up with an empty configuration file. Edit the file to your liking, save and close the application.

CHAPTER 4

Supported input formats

4.1 Compendex

Todo

4.2 Scopus

Todo

4.3 Web of Science

Todo

CHAPTER 5

Generating graphs

5.1 Graph types

Currently three types of graphs are supported:

1. Country
2. Institution
3. Author

These graphs types can be seen as three differently grained levels of the same dataset. The first one is the most coarse grained, while the last one is the most fine grained.

5.2 Filtering publications

5.2.1 Filtering by publication year

It is possible to limit the result set for which graphs are generated to publications occurred during certain years.

To activate the filtering, it suffices to pass a value for the `--years` option when invoking the `irco-graph` command.

The `--years` options can parse the following values:

- A single year: `2012`
- A list of single years: `2003,2007,2013`
- A range of years (inclusive on both ends): `2003-2006` (equivalent to `2003,2004,2005,2006`). Additionally a range can be open on one of its end, in which case no limiting will occur on that side:
  - `2012-` will select all publications with a publication date of 2012 or later;
  - `-2002` will select all publication with a publication date of 2002 or earlier.
- A combination of single years and ranges: `2003-2006,2009,2012-`

The following command creates a `country` graph with all papers published in or before 2000, in 2002, in 2003, in 2004, in 2005, in 2006, in 2008, in 2009, or after 2013 (included):

```
irco-graph --years 2008,2009,2002-2006,-2000,2013- country sqlite:///test.db test.gexf
```

5.2.2 Filtering by corresponding author country

Publications can be filtered by ore more corresponding author countries. When active, this filter will only show publications for which the country of the institution of the corresponding author is in the list of defined values.

To activate the filtering, it suffices to pass one or more values for the `--ca-country` option when invoking the `irco-graph` command. The option can be repeated multiple times to specify more than one alowed countries.

The short hand version of the argument, `-c`, can be used as well and the country name matching is case insensitive.

The following command creates a `country` graph with all papers which have a corresponding authors affiliated to an institution residing in either Kuwait or Qatar:

```
irco-graph --ca-country=Kuwait -c qatar country sqlite:///test.db test.gexf
```

5.2.3 Filtering by publication type

Publications can be filtered by their type. The currently known types are: `journal`, `conference`, `book`, `book in series` and `patent`.

Filtering by publication type works similarly as with the corresponding author country filter, but by using the `--type` (or `-t` short hand) command line option.

The followign command creates a `country` graph with all journal articles or book publications:

```
irco-graph --type=book -t journal country sqlite:///test.db test.gexf
```

5.3 Misc notes

5.3.1 Problems with the *Institution* graph

The current implementation of the *Institution* graph takes the institution name as the key to create graph nodes. This behaviour induces the system to create numerous nodes for the same entity as the instituion name is not normalized in the data sets from which the database is populated.

For example, in one of the examined testing data sets, the *"Carnegie Mellon University"* appears in at least 19 different variations of its name:

```
Carnegie Mellon Qatar, Qatar
Carnegie Mellon University - Qatar, Doha, Qatar
Carnegie Mellon University In Qatar, P.O. Box 24866, Doha, Qatar
Carnegie Mellon University in Qatar, Compute Science Department, Doha, Qatar
Carnegie Mellon University in Qatar, Doha, Qatar
Carnegie Mellon University in Qatar, Education City, Doha, Qatar
Carnegie Mellon University in Qatar, Education City, PO Box 24866, Doha, Qatar
Carnegie Mellon University in Qatar, P.O. Box 24866, Doha, Qatar
Carnegie Mellon University in Qatar, PO Box 24866, Doha, Qatar
Carnegie Mellon University in Qatar, Qatar Cloud Computing Center, Qatar
Carnegie Mellon University in Qatar, Qatar Foundation, Education City, P.O. Box 24866, Doha, Qatar
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States
Carnegie Mellon University, Doha, Qatar
Carnegie Mellon University, Education City, PO Box 24866, Doha, Qatar
Carnegie Mellon University, Heinz College, Pittsburgh, PA, United States, Qatar Campus, Doha, Qatar
Carnegie Mellon University, P.O. Box 24866, Qatar, Qatar
Carnegie Mellon University, Pittsburgh, PA, United States
Carnegie Mellon University, Qatar
```

```
Carnegie Mellon University, Qatar Campus, PO Box 24866, Doha, Qatar
Carnegie Mellon University, Qatar Education City, Doha, Qatar
```

Also note that this university exists once with its original name in Pennsylvania and as a branch campus in Qatar (with the *"in Qatar"* suffix).

Different approaches can help solve (or at least reduce) the impact of this problem:

1. Use of a normalized data set
2. Normalization of the data set with data mining techniques

In the second case (*in-house normalization*), the following non-exhaustive list of techniques can be employed:

1. Normalization through geolocation
2. Normalization through text analysis (pattern matching + similarity measures)
3. Exploitation of other information respositories (Google searches, Wikipedia articles, ...)
4. Manual matching (crowdsourcing, Amazon Machine Turk, ...)

CHAPTER 6

Command line utilities reference

The IRCO packages comes with different command line utilities to work on data sets.

6.1 irco-convert

Deprecated.

6.2 irco-explorer

Starts the IRCO Explorer server.

6.3 irco-geolocate

TODO: This command is not yet ready.

6.4 irco-graph

Generate a graph file from a dataset.

6.5 irco-import

Import a data set into an IRCO database.

6.6 irco-init

Initializes a new IRCO database.

CHAPTER 7

Downloading search results from WoS

7.1 Shell script

IRCO version >= 0.9 provide the *irco-scrape* command to download search results from the Web of Science database. The usage of the command is simple:

```
irco-scrape <search-id> path/to/output/directory
```

Where `<search-id>` has to be replaced with the WoS search id, as found in the URL (under the named GET parameter `SID`).

7.2 Bookmarklet

To simplify the construction of the command, along with its correct `<search-id>` parameter, you can add the link below to the bookmarks of your browser (normally it suffices to drag & drop it to the bookmarks bar):

You can then open it while displaying the search results page to get a popup window containing the command to run. Note that you still have to copy & paste the command into a shell and run it.

Todo

This command currenly just downloads the files to the given folder. You still have to manually import them using the `irco-import` command.

CHAPTER 8

Backing up the database

When working on the whole data set using possibly destructive IRCO commands, it is strongly suggested to back up the database to be able to restore the data in case something goes wrong.

As IRCO does not provide a built-in, cross-platform utility to back up the data easily, the following page describes how such a backup can be created for specific database management systems.

8.1 SQLite 3

SQLite 3 provides an easy to use tool to backup a database. To create the backup, issue the following command:

```
sqlite3 path/to/the/database.db .dump > path/to/the/backup.bak
```

If you ever need to restore the data from a backup, use the following command:

```
mv path/to/the/database.db path/to/the/database.db.old
sqlite3 path/to/the/database.db < path/to/the/backup.bak
```

Additional references:

- http://www.sqlite.org/backup.html
- http://www.ibiblio.org/elemental/howto/sqlite-backup.html

CHAPTER 9

News

9.1 Release notes

9.1.1 News for IRCO 0.6

Upgrade

You can install the latest version of the irco tool by issuing the following command:

```
pip install --upgrade irco
```

At the time of writing the latest version is 0.6. There are some major news for this release, as better described below.

Database

The new database system is in place. This will complicate things initially, but will be a good choice for the future.

I have some more work to do before giving you access to a centralized database. For the moment you can use a local SQLite database. This means that in all places where a database connection string is required, you can use the following:

```
sqlite:///<name-of-the-database.db>
```

or:

```
sqlite:///</absolute/path/to/name-of-the-database.db>
```

In the first case, the path is relative to your working directory, while in the second, it is absolute to the root of the hard disk. You can place the database wherever you want and even using more than one database (e.g., to keep records separated). I think that a sane default is:

```
sqlite:///irco.db
```

Before running any other command, you have to initialize the database:

```
irco-init sqlite:///irco.db
```

Then you can import files from different sources:

```
irco-import -i scopus scopus.csv sqlite:///irco.db
```

or, to import a file in compendex format:

```
irco-import -i compendex compendex.txt sqlite:///irco.db
```

When you have imported a bunch of files, you can still generate the authors or country graphs by replacing the path to the source file with the database:

```
irco-graph authors sqlite:///irco.db out.gexf
```

or, to generate the *country* graph:

```
irco-graph country sqlite:///irco.db out.gexf
```

The old `irco-convert` command is deprecated and should not be used anymore.

Documentation

I started to write the documentation for the tool. It does not contain anything yet (except this page), but that's the next task on my todo list.

You can always find it at this address: http://irco.readthedocs.org/

IRCO Explorer

IRCO Explorer is an interactive record explorer that allows to browse the database from a web browser. You can run it locally by issuing this command:

```
irco-explorer sqlite:///irco.db
```

and then navigating to the following page from you preferred web browser:

> http://localhost:8000/

When you are done exploring the dataset, hitting `Control + C` in the terminal windows where you executed the `irco-explorer` command terminates the server.

9.1.2 News for IRCO 0.6.1

Upgrade

You can install the latest version of the irco tool by issuing the following command:

```
pip install --upgrade irco
```

At the time of writing the latest version is 0.6.1. This is a minor release which introduces some user interface related changes, as better described below.

Explorer update

Version 0.6.1 is a small update to the explorer which adds styling and pagination in order to provide a faster and overall more enjoyable user experience.

For more information about the *IRCO Explorer*, see *IRCO Explorer*.

Next steps

The next version will provide additional utilities to work on data, such as institution geolocation and entries deduplication (merging of institution or author records refering to the same entity).

9.1.3 News for IRCO 0.9

Upgrade

You can install the latest version of the irco tool by issuing the following command:

```
pip install --upgrade irco
```

Changelog

- `irco-scrape` command to download WoS search results.
- Records with ambiguous author affiliations are ignored. The number of ignored records due to ambiguity is reported at the end of each `irco-import` as `Records with ambiguous auth. aff.`.
- The affiliation in the *reprint author* field (`RP`) takes precedence over the affiliation listed in the *affiliations* field (`AF`).

CHAPTER 10

Indices and tables

- *genindex*
- *modindex*
- *search*