CSG Dicoms Anonymizer

Stephen Larroque
Coma Science Group, GIGA research
University of Liège

07/04/2017

Advantages of CSG Dicoms Anonymizer

- **Automatically** detect name and **remove** it from any field, even **hidden ones**
- **Uniformize names** (invariant to typos & words switching)
- Can anonymize **demographics** along (same anon ids)
- Can use same demographics file for any set of dicom –> anonymization will shorten to pertinent subjects
- Can continue a partial anonymization (eg, bug, access denied, …)
- Deterministic anonymization → Can **update** anonymized demographics
- Generate list of **missing** demographics (ie, dicom files are present but no demographics for them).
- Generate a set of csv files to **deanonymize**.
  Tip: These files can be encrypted in a 7z (not zip) archive with password and sent to the collaborator, he'll send it back if need more infos (ie, less work and files storage for us).

Dicoms Anonymizer – Algorithm

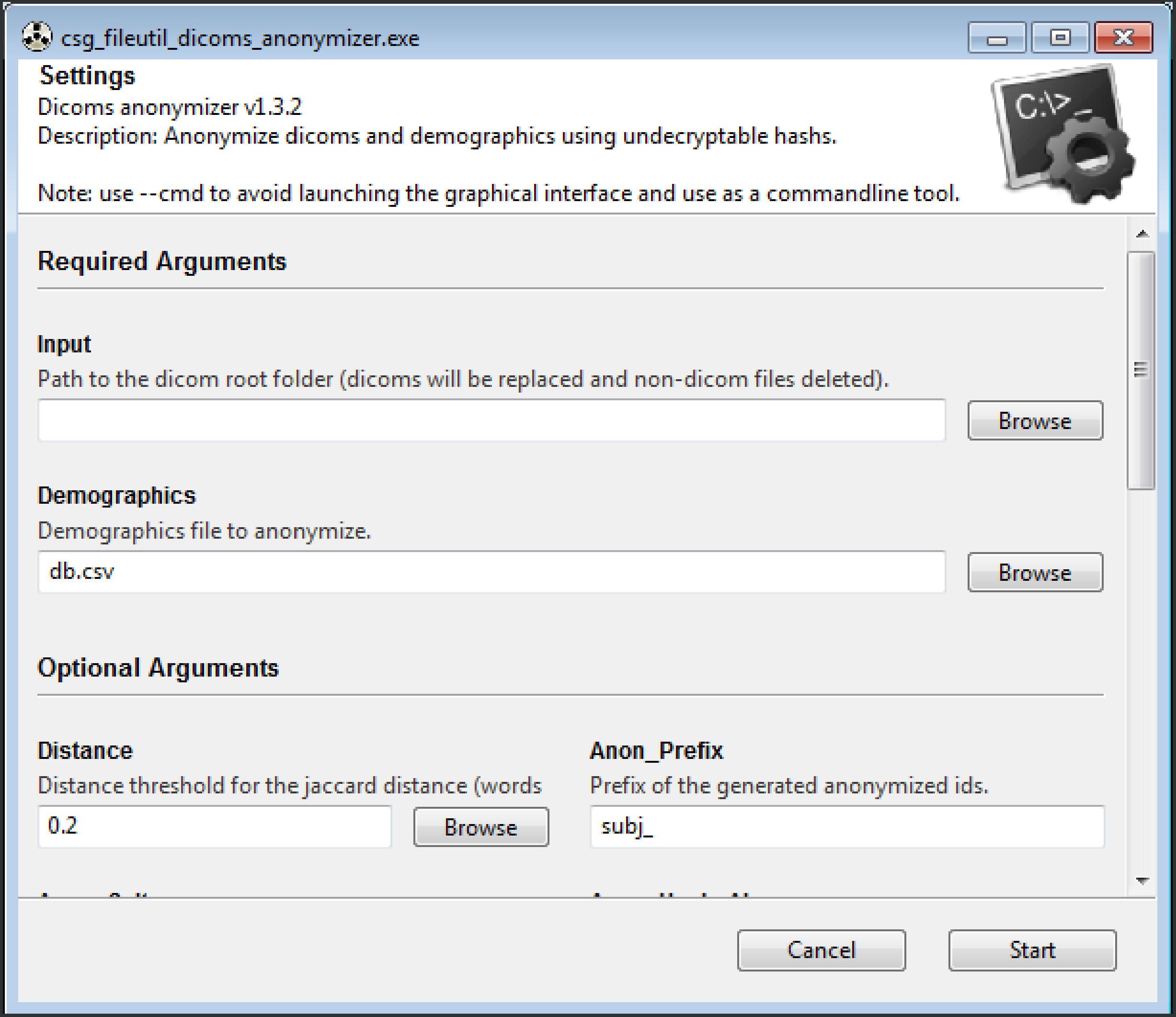
1. Generate list of patient names from dicoms (folders and zips)
2. Generate unique list of names (disambiguate similar names)
→ dicom_names.csv
3. Generate MD5 hash from names (with salt if provided → each lab can generate unique deterministic ids by tweaking the salt)
4. anonymized id = Shortened MD5 hash
→ idtonames.csv
5. If demographics: merge with dicoms names (compute distance matrix using disambiguation based on letters + words normalized levenshtein distance)
6. Apply anonymized id to dicoms files, folders names and demographics → anonymized dicoms & demographics
7. Delete non dicom files (pdf, doc, docx, txt, etc.)

Dicoms Anonymizer – Usage

1. Copy **all dicom** folders/zips in one folder
2. Get **demographics** file (optional)
3. Open **csg-fileutil dicoms anonymizer** (using Jupyter Notebook).
4. **Replace parameters** (dicom rootpath, demographics file path) **under each Part x**.
5. Click **Kernel > Restart & Run All**

→ Result: anonymized dicoms & demographics

Dicoms Anonymization – GUI



Settings

Dicoms anonymizer v1.3.2

Description: Anonymize dicoms and demographics using undecryptable hashs.

Note: use --cmd to avoid launching the graphical interface and use as a commandline tool.

Required Arguments

Input

Path to the dicom root folder (dicoms will be replaced and non-dicom files deleted).

Demographics

Demographics file to anonymize.

db.csv

Optional Arguments

Distance

Distance threshold for the jaccard distance (words

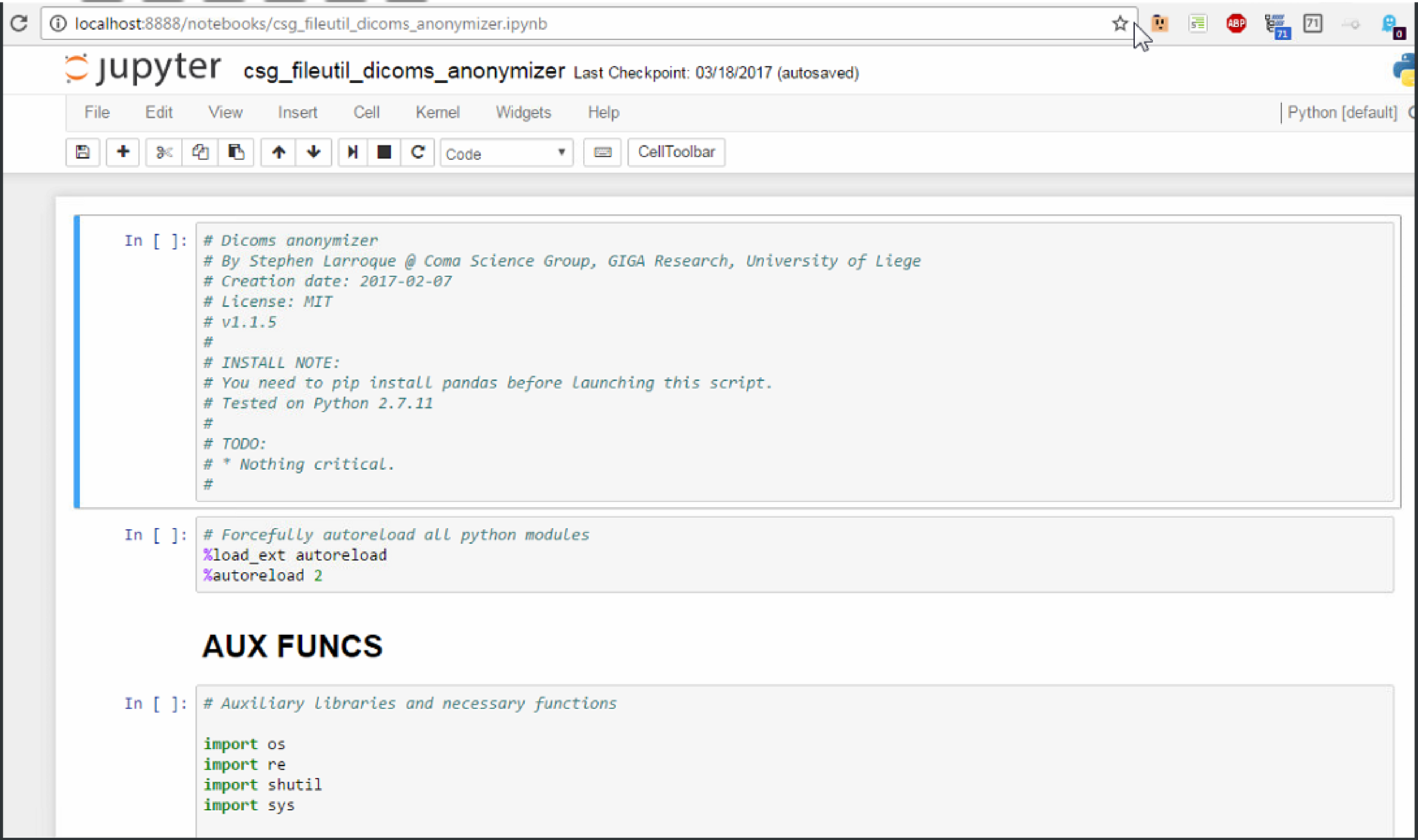
0.2

Anon_Prefix

Prefix of the generated anonymized ids.

subj_

Dicoms Anonymization – Jupyter Notebook (old, please use GUI now)


localhost:8888/notebooks/csg_fileutil_dicoms_anonymizer.ipynb
jupyter
csg_fileutil_dicoms_anonymizer Last Checkpoint: 03/18/2017 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python [default]
Code
CellToolbar

In [ ]:

```
# Dicoms anonymizer
# By Stephen Larroque @ Coma Science Group, GIGA Research, University of Liege
# Creation date: 2017-02-07
# License: MIT
# v1.1.5
#
# INSTALL NOTE:
# You need to pip install pandas before launching this script.
# Tested on Python 2.7.11
#
# TODO:
# * Nothing critical.
#
```

In [ ]:

```
# Forcefully autoreload all python modules
%load_ext autoreload
%autoreload 2
```

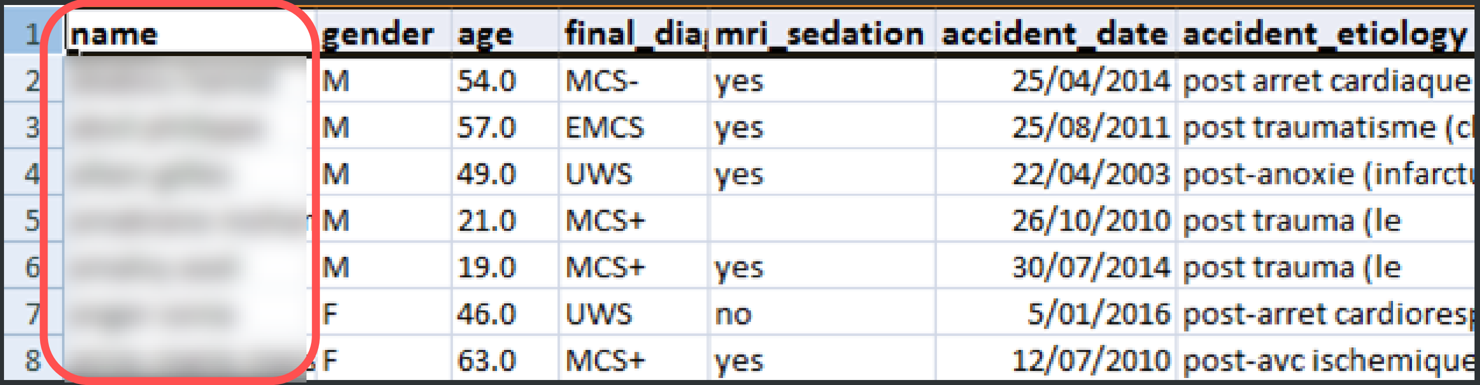
AUX FUNCS

In [ ]:

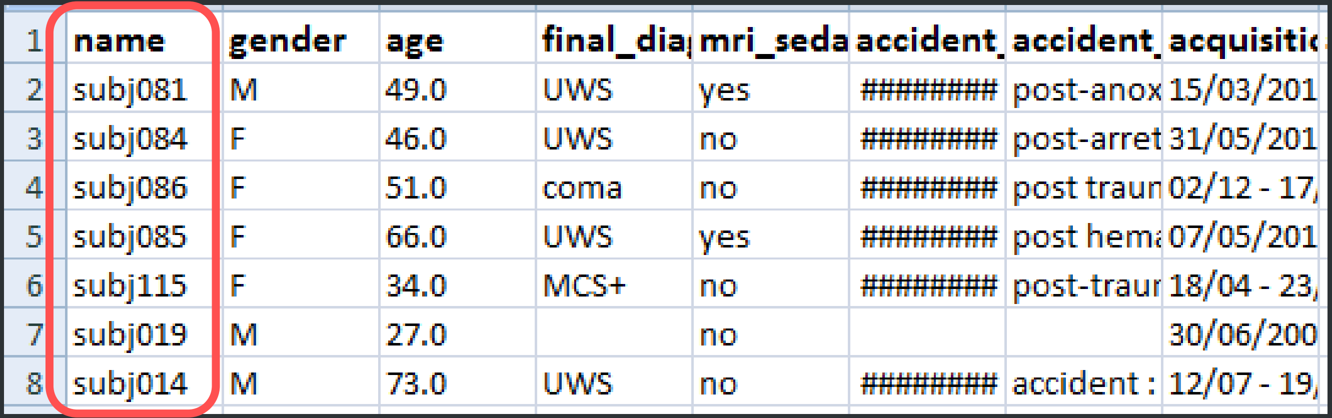
```
# Auxiliary libraries and necessary functions

import os
import re
import shutil
import sys
```

Anonymized demographics



	name	gender	age	final_dia	mri_sedation	accident_date	accident_etiology
2	[redacted]	M	54.0	MCS-	yes	25/04/2014	post arret cardiaque
3	[redacted]	M	57.0	EMCS	yes	25/08/2011	post traumatisme (cl
4	[redacted]	M	49.0	UWS	yes	22/04/2003	post-anoxie (infarctu
5	[redacted]	M	21.0	MCS+		26/10/2010	post trauma (le
6	[redacted]	M	19.0	MCS+	yes	30/07/2014	post trauma (le
7	[redacted]	F	46.0	UWS	no	5/01/2016	post-arret cardioresp
8	[redacted]	F	63.0	MCS+	yes	12/07/2010	post-avc ischemique



	name	gender	age	final_dia	mri_seda	accident_	accident_	acquisitic
2	subj081	M	49.0	UWS	yes	#########	post-anox	15/03/201
3	subj084	F	46.0	UWS	no	#########	post-arret	31/05/201
4	subj086	F	51.0	coma	no	#########	post traun	02/12 - 17,
5	subj085	F	66.0	UWS	yes	#########	post hema	07/05/201
6	subj115	F	34.0	MCS+	no	#########	post-traur	18/04 - 23,
7	subj019	M	27.0		no			30/06/200
8	subj014	M	73.0	UWS	no	#########	accident :	12/07 - 19,

Resulting files

What you can send:

□ Anonymized dicoms

□ Anonymized demographics (demographics_anonymized_shortened.csv)

□ Missing demographics anonymized (missing_demo_anonymized.csv)

What you need to keep (but not send):

□ **idtoname.csv → anonymization mapping**, if collaborator might need more info about 1 subject

□ dicom_names.csv → to regen anon (eg, to update demo)

□ missing_demo.csv → missing demographics

□ demographics_shortened.csv (optional)

Dicom Anonymization – Tips & tricks

- □ **Dicoms will be replaced**, advised to **backup (zip)** before anonymization (in case something went wrong and you need to restart)
- □ Anonymization can be continued if error or stopped (but disadvised)
- □ Demographics automatically **shortened** to subjects present in dicoms → can use the same demographics for all anonymizations
- □ Script divided in **3 independent parts**: 1. extract dicom names, 2. generate anonymization mapping, 3. anonymize dicoms & demographics.
  → **Can restart at any part (eg, to update demo)**

Thank you for your attention

COMA SCIENCE GROUP
GIGA
fnrs FREEDOM TO RESEARCH
.be
James S. McDonnell Foundation
MIND SCIENCE FOUNDATION
CHU de Liège
Université de Liège
ULg

BONUS SLIDES

Université de Liège

G I G A

Dicoms Anonymizer – Old Algorithm

1. Generate list of patient names from dicoms (folders and zips)
2. Generate unique list of names (disambiguate similar names)
   → dicom_names.csv
3. Generate MD5 hash from names
4. Reorder names by MD5 hash
5. New order = anonymized id
   → idtonames.csv
6. If demographics: merge with dicoms names (compute distance matrix using disambiguation based on letters + words normalized levenshtein distance)
7. Apply anonymized id to dicoms files, folders names and demographics → anonymized dicoms & demographics
8. Delete non dicom files (pdf, doc, docx, txt, etc.)

□ Future: add salt for anonymized id (so that each lab can generate its own unique deterministic ids)