OGD in research iConference version documentation

An Yan, 10.20. 2017

# Open data sites selection

Original : <https://www.data.gov/open-gov/>

Parse portal names to strip “http” and “www”

* input: 'search\_terms.csv' which is the original portals downloaded from data.gov
* Code: parse\_OGD\_portalnames.ipynb
* Output: trimmed portal urls in strings.
* Manual edits: check if urls still work, if not, replace with new urls, or remove them

springer\_search\_string.txt, IEEE\_scopus\_search\_string.txt

(these are the strings put into Database searchers.)

# Database selection and article retrieval

Time: 20170801

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Scopus | IEEE | Springer | total |
| Init retrieve | 1997 | 263 | 226 | 2486 |
| Removed dup | 1997 | 177 | 202 | 2376 |

* Retrieving papers

**in Scopus:**

ALL ( "odaa.dk" OR "open.alabama.gov" OR "data.alberta.ca" OR "cabq.gov/abq-data" OR "alkmaar.nl/opendat" OR "datacatalogs.org/catalog/allerdal" OR "os.amsterdam.nl" OR "data.angers.fr" OR "a2gov.org/data" OR "opendata.antwerpen.be" ) AND PUBYEAR > 2008 AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) )

**in IEEE:**

("odaa.dk" OR "open.alabama.gov" OR "data.alberta.ca" OR "cabq.gov/abq-data" OR "alkmaar.nl/opendat" OR "datacatalogs.org/catalog/allerdal" OR "os.amsterdam.nl" OR "data.angers.fr" OR "a2gov.org/data" OR "opendata.antwerpen.be") and refined by Content Type: Conference Publications Journals & Magazines Early Access Articles   Year: 2009-2018

notes: when retrieving 'data.gov' in Scopus, additional conditions used: excluded papers which has 'open government data' in title/keywords/abstract. Otherwise, there are way too many results...

**Remove duplicates:**

* Code: IEEE\_springe\_clean.ipynb
* Input: 'IEEE\_total\_20170801.csv', Springer\_total\_20170801.csv, scopus\_total\_20170801.csv
* Output: Springer\_total\_20170821\_cleaned.csv, IEEE\_total\_20170821\_cleaned.csv

# sample annotation

**Input**: Springer\_total\_20170821\_cleaned.csv,

IEEE\_total\_20170821\_cleaned.csv,

scopus\_total\_20170801.csv

**results summary**:

In/out: sample size: 170

Usage type 1 (round 2+ round 3 + round 4): generate kappa, additional use types were compared as well. So in total there are 29+ 24 + 10 = 63 records, but plus additional types, there are in total 82 records in kappa sheet.

Usage type2 (round 5). There are 49 samples, plus additional types, there are 51 records.

**output**: samples\_annotate folder

kappa-calculator-OGD-inout\_20170828.xlsx

kappa-calculator-usagetypes1\_20170910

kappa-calculator-usagetypes2\_20170910

**procedure**:

* retrieve materials and remove duplicates (August, 4th)
* Nic and Annie examine 50 samples (From Scopus sample 1 ~ 50) to determine in/out (August, 4th)
* Nic and Annie examine anther 98 samples (Springer 1~98) to determine use types (August, 11st)
* Nic and Annie examine anther 52 samples (Springer 99~150) to determine in/out and use types (August, 18st)
* Nic and Annie examine 10 samples (Springer 185, 191-194, 196, 198,199,200,203) to determine usage types (8.29)
* Nic and Annie examine 50 samples (Scopus, 52-114, only those determined as "in") to determine usage types (8.29)

**Annie's annotation of first 50 samples:**<https://docs.google.com/a/uw.edu/spreadsheets/d/1DrfMxdu4L1FjeGx9T7e3JMDp5APZ52lj40awMphG4iU/edit?usp=sharing>

Nic and Annie determine in this round:

* whether an article should be included in our analysis according to Inclusion/Exclusion criteria
* what purposes OGD serve in a research paper, see <https://github.com/OpenDataLiteracy/ODG-usage-in-Research/wiki/OGD-use-cases-and-usage-types> for types of use and examples.

**Annie's annotation of second round 100 samples from Springer:**

<https://docs.google.com/a/uw.edu/spreadsheets/d/1PufdOkukrXerqiuvrhhSv1aFcw1hcU016UvYDvP2H58/edit?usp=sharing>

or use copy on Github : <https://github.com/OpenDataLiteracy/ODG-usage-in-Research/blob/master/lit_mining/Springer_sample2_20170808_Annie.xlsx>

Nic and Annie determine in this round:

* what purposes OGD serve in a research paper, see <https://github.com/OpenDataLiteracy/ODG-usage-in-Research/wiki/OGD-use-cases-and-usage-types> for types of use and examples.

**Third round of annotation of 50 samples from Springer**

* Google spreadsheet : <https://docs.google.com/a/uw.edu/spreadsheets/d/1AX7C-R4QXs-apq8X2U_h1VLANfLvLyIb0Y571ZcwmZo/edit?usp=sharing>
* Github link: <https://github.com/OpenDataLiteracy/ODG-usage-in-Research/tree/master/lit_mining>

Nic and Annie determine in this round:

* whether an article should be included in our analysis according to Inclusion/Exclusion criteria
* what purposes OGD serve in a research paper, see <https://github.com/OpenDataLiteracy/ODG-usage-in-Research/wiki/OGD-use-cases-and-usage-types> for types of use and examples.

**4th round of annotation, 10 samples:**<https://docs.google.com/a/uw.edu/spreadsheets/d/1MX2FjN1HyJ3x4RCAPdDmqMeFxsvMTR3loOSRZfpZKBg/edit?usp=sharing>

\*\*5th round of annotation, 50 samples: \*\*<https://docs.google.com/a/uw.edu/spreadsheets/d/10Om6eWpjNqiCTJbs1pttKZxJlMKVKXvCHxu8vzfLqYg/edit?usp=sharing>

# Annotation

One researcher annotated rest of all samples. Determined usage type and in/out

**Input**: Springer\_total\_20170821\_cleaned.csv,

IEEE\_total\_20170821\_cleaned.csv,

scopus\_total\_20170801.csv

**output:**

Usage types (after removing papers that were out, IEEE = 36, scopus = 1080, springer = 80, total=1194)

Note: when parsing pdf to txt, there are some files lost, when extract url from txt, there are other files lost in Scopus. So after extracting urls, there were 1140 papers remained. The 1194-1140 = 54 gap was left unaddressed.

There are papers belong to more than one usage types, so there are 58 additional usage types. Additional types : 58 + 1194=1252

* **fulltext for all papers that are included, see folders under folder “annotation\_round 1”**

fulltext\_IEEE, fulltext\_scopus, fulltext\_springer

* **annotated list of papers:**

IEEE\_total\_Annie\_20170811.xlsx

scopus\_total\_51\_1997\_Annie.xlsx

Springer\_total\_Annie\_20170808.xlsx

combined\_use\_types.csv

# Parse fulltext pdf to txt

Convert pdf fulltext files to .txt formats for processing.

* Code: parse\_pdf.ipynb
* Input:

springer\_search\_string.txt, IEEE\_scopus\_search\_string.txt

fulltext\_IEEE, fulltext\_scopus, fulltext\_springer

* Output:

Springer\_txt, ieee\_txt, scopus\_txt

* Manual: fixed broken files and duplicated file names

Note: fulltext pdf files and txt files were not uploaded to Github due to space restrictions.

* Summary: resulting scopus 1023 + springer 82 + IEEE 35, Total: 1140 as unique papers included). Papers are citing 98 out of 302 unique OGD portals.

# extract URL from full text txt files

Extract OGD portal urls from txt files. There are 1140 papers, but papers can have multiple citations, so there are 1189 unique URLS extracted.

* Code: extract\_url.ipynb, extract\_url\_springer.ipynb

**Step 1: clean search strings**

* Input 1:

springer\_search\_string.txt, IEEE\_scopus\_search\_string.txt

Search.term.csv

* Output 1:

'merged\_search\_string.csv’ (for IEEE and scopus), springer\_merged\_search\_string.csv

* Manual edits: added Taiwan and Israel OGD portals and cleaned.
* Output from manual edits:

search\_string\_IEEE\_edited2.csv (for IEEE and scopus)

search\_string\_springer\_edited.csv

**Step 2: use search strings to extract urls from txt files.**

* Input 2:

Springer\_txt, ieee\_txt, scopus\_txt

* Output 2:

extract\_url\_IEEE.csv, scopus\_extract\_url\_df.csv, springer\_extract\_url\_df.csv

* Manual cleaning: removed empty records

combined\_extract\_url\_cleaned.csv

scopus\_extract\_url\_cleaned.csv

extract\_url\_IEEE\_cleaned.csv

springer\_extract\_url\_cleaned.csv

# Producing stat and plotting

* Code: extract\_url\_springer.ipynb
* Input 1 : combined\_extract\_url\_cleaned.csv
* Output 1: stat\_plot.xlsx (Top 30 sources of OGD used in scientiﬁc research)
* Input 2 : combined\_use\_types.csv
* Output 2: stat\_use\_types.xlsx (year breakdown, use types breakdown)
* Input 3 : Scopus\_Source\_List\_used.csv

Note: downloaded from Scopus content coverage website:

<https://www.elsevier.com/solutions/scopus/content>

* Output 3: title\_to\_areas.csv
* Manual edits: match publications to scopus areas:

title\_to\_areas\_manual\_final.csv

merged\_types\_areas.csv

code\_count.csv (breakdown of research areas)