

从元数据到RDF数据

OpenRefine作为RDF数据转换工具

上海图书馆 夏翠娟
2018年6月 @****大学

主要内容

- OpenRefine简介——使用范围和场景
- OpenRefine下载安装
- OpenRefine数据导入
- OpenRefine数据清洗
- OpenRefine的RDF相关模块的使用
- 知识本体映射配置
- RDF数据导出
- 多源数据混搭

OpenRefine简介

- Google Refine2.5
- OpenRefine 2.6 (Oct 13, 2015)
- 数据剖析 (data profiling)
- 数据清洗 (data cleaning)
- 数据转换 (interactive data transformation tools)
- 数据关联与混搭
- 看起来像EXCEL, 用起来像database
- 不方便增加新的行

OpenRefine 下载

- 下载: <http://openrefine.org/download.html>
- 版本
 - OpenRefine 2.6-rc2 Release Candidate 2
 - OpenRefine 2.6 beta 1
 - Google Refine 2.5
 - [LODRefine](#)
- 插件
 - [RDF extension 0.8](#) for Google Refine 2.5
 - [RDF extension 0.9](#) for OpenRefine 2.6

OpenRefine安装

•OpenRefine在Windows的安装

- 1) ZIP包解压到某个目录;
- 2) 要运行OpenRefine, 双击openrefine.exe文件。

•OpenRefine在Mac的安装

- 1) DMG文件打开磁盘镜像, 拖动OpenRefine的图标到Applications目录;
- 2) 双击图标以启动OpenRefine。

OpenRefine在Linux的安装

- 1) gzipped包解压到当前用户的home目录;
- 2) 在终端命令行环境, 键入./refine以启动OpenRefine。

OpenRefine 安装 RDF 插件

解压到\google-refine-2.5-r2407\webapp\extensions\rdf-extension

Google refine 世系表 Permalink

Facet / Filter Undo / Redo 4

42 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: Fr...base ▾ RDF ▾

« first < previous - 10 next > last »

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

	代	personID	谱名	名	字	号	排行	生	卒	迁徙	说明
1.	42	16610	洪驊	适	适之		三	光绪辛卯十一月十七未时			考派美国留学生名适字适之事迹见学了生光绪辛卯十一月十七未时娶江氏生光绪庚寅十一月初八辰时
2.	41	41	祥蛟	传	守三	铁花、钝夫	一	道光辛丑二月十九戌时	光绪乙未七月初三子时		岁贡生诰授通议大夫钦加三品衔戴花翎台湾台湾直隶洲知州核领镇海后军在任候补知府事迹分见仕宦学林善行
3.	40	40	贞铸	奎熙	世恩	律均	二	道光壬午三月初九辰时			诰封奉政大夫晋封通议大夫娶程氏生嘉庆庚辰七月十六时卒光绪丙戌二月十一申时生五子则室张氏生道光壬寅十一月初五
4.	39	39	锡镛		序东		一	乾隆壬子十二月初十	道光壬午三月初二十三		诰封奉政大夫晋封通议大夫娶曹氏节行见列女生乾隆癸丑六月十八卒同治壬戌三月初十生二子合彝余村岱
5.	38	38	德江	瑞杰	宗海		二	乾隆丁亥三月十八	道光甲辰九月十八		国学生诰封通议大夫娶曹氏生一子继娶曹氏生乾隆丁未卒咸丰丙辰生二子公与继配曹氏合葬黄泥岭山向详载墓图元配曹氏葬洪家
6.	37	37	天旭	旭	日初		一	乾隆甲子三月二十二			国学生事迹见善行娶程氏生乾隆甲子三月二十二生三子合彝余村岱

OpenRefine 导入数据

« Start Over
Configure Parsing Options
Project name: 世系表20161210
Create Project »

Create Project

Open Project

Import Project

代	personID	father	谱名	名字	号	排行	生	卒	迁徙	说明	
1.	40	33141	33132	贞焘	奎照	崇恩	星五	一	嘉庆甲戌八月二十四	同治乙丑正月初九	邑庠优增生事迹见学林 娶曹氏生嘉庆辛未 继娶石氏生嘉庆辛巳二月初二 歿咸丰辛酉八月初十 生二子三 娶汪氏 生道光庚寅六月十九 歿宣统己酉十一月十三 曹氏葬刘家塘石氏葬蟹形
2.	41	33140	33133	祥麟	玮	守玉	汉生	二	道光丙午二月初二	光緒己亥	国学生娶曹氏贤行 见列女生道光戊申十一月十一 生一子
3.	41	33139	33133	祥系	巳	守璋		三	道光戊申正月十八辰时	光緒戊子二月十八午时	娶姚氏生道光庚午三月十五丑时 生二子
4.	41	33138	33133	祥纘	玠	守燥	介如	四	咸丰甲寅正月十五	光緒甲辰十月十七	廩贡生阜阳县训导事迹见 见仕宦学林 娶朱氏生咸丰丁巳八月十六 卒光緒辛卯十二月初一 生二子 继娶林氏 二子以次子洪驥为祥麟后
5.	41	33137	33133	祥祐	守翊		吉庭	五	咸丰丙辰元旦午时		翰林院待诏娶汪氏生咸丰丁巳七月十六酉时
6.	42	33136	33134	洪骏	嗣稼	耕云		一	同治辛未正月十四辰时		从九品娶章氏生同治己巳十一月十三卯时 生二子
7.	42	33135	33134	洪骏	国琦	绍之		二	光緒丁丑八月二十三戌时		国学生候选知县娶汪氏生光緒戊寅十二月十一申时 生二子
8.	42	16610	33134	洪骏	适	适之		三	光緒辛卯十一月十七未时		考派美国留学生名适字适之事 见学了生光緒辛卯十一月十七未时 娶江氏生光緒庚寅十一月初八辰时
9.	41	33134	33133	祥蛟	传	守三	铁花、钝夫	一	道光辛丑二月十九戌时	光緒乙未七月初三子时	岁贡生诤授 通议大夫 欽加三品 銜 戴花翎 台湾台北直隶州知州 统辖镇海 后军在任 候补知府 事迹 见仕宦学林 善行
10.	40	33133	33132	贞翰	奎熙	世恩	律均	二	道光壬午三月初九		诤封奉政大夫 晋封 通议大夫 娶程氏生嘉庆庚辰七月十六 时卒光緒丙戌二月十一申时 生五子 侧室张氏生道光壬寅十一月初五

Parse data as

Excel files

JSON files

Line-based text files

CSV / TSV / separator-based files

Fixed-width field text files

PC-Axis text files

RDF/N3 files

XML files

Worksheets to Import

世系表 50 rows

Ignore first 0 line(s) at beginning of file

Parse next 1 line(s) as column headers

Discard initial 0 row(s) of data

Store blank rows

Store blank cells as nulls

Store file source (file names, URLs) in each row

Update Preview

Version 2.5 [r2407]

Help

OpenRefine数据清洗

[GREL \(Google Refine Expression Language\)](#)

- [Variables](#)
- [GREL Controls](#)
- [GREL Functions](#) overview
- [GREL Boolean Functions](#)
- [GREL String functions](#), including parsing, splitting, encoding and hashing
- [GREL Array functions](#)
- [GREL Math functions](#)
- [GREL Date functions](#)
- [GREL Other functions](#) including JSON and Jsoup

OpenRefine数据清洗

1. 数据类型转换

```
substring(toString(value),0,indexOf(toString(value),"."))
```

2. 生成新的列 (Edit column->add column based on this column)

```
substring(value,0,2)
```

```
if(indexOf(value,"娶")>-1,substring(value,indexOf(value,"娶")+1,indexOf(value,"氏")+1),"")
```

3. 分面浏览

4. 批量修改

OpenRefine数据转换

1. 设置Base URI
2. 设置命名空间前缀
3. 设置“主体”的类型
4. 设置“主体”的URI规则
5. 设置“谓词”及元数据字段与本体属性之间的映射
6. 设置“客体”的取值
7. 随时查看转换结果 (RDF Preview)
8. 导出RDF数据

来自本体词表中的“谓词”

属性名 (Name)	标签 (Label)	subPropertyOf	域 (Domain)	范围 (Range)
rdfs:label	标签		rdf:Resource	Literal
c:identifier	标识符		shl:Resource	Literal
rl:description	描述		shl:Resource	Literal
rl:temporal	朝代		shl:Resource	shl:Temporal
c:date	日期		shl:Resource	Literal
rl:event	事件		shl:Resource	shl:Event
rl:place	地点		shl:Resource	shl:Place
oaf:name	名称		rdf:Resource	Literal
oaf:gender	性别		shl:Person	bf:Category
oaf:familyName	姓氏		shl:Person	shl:FamilyName
rl:genealogyName	谱名		shl:Person	Literal
rl:name	名称		shl:Agent	shl:Name
rl:courtesyName	字		shl:Person	Literal
rl:pseudonym	号		shl:Person	Literal
rl:orderOfSeniority	行		shl:Person	Literal
rl:posthumousTitle	谥号		shl:Person	Literal
rl:birthday	生于		shl:Person	Literal
rl:deathday	卒于		shl:Person	Literal
rl:birthPlace	出生地		shl:Person	shl:Place
rl:deathPlace	死亡地		shl:Person	shl:Place
rl:ethnicity	民族		shl:Person	bf:Category
rl:nationality	国籍		shl:Person	bf:Category
rl:nativePlace	籍贯		shl:Person	shl:Place
rl:speciality	专长		shl:Person	Literal
rl:briefBiography	小传		shl:Person	Literal
rl:createdWork	著作		shl:Person	Literal

属性名 (Name)	标签 (Label)	subPropertyOf	域 (Domain)	范围 (Range)
hl:briefBiography	小传		shl:Person	Literal
hl:createdWork	著作		shl:Person	Literal
hl:officialExperience	履历		shl:Person	Literal
hl:officialEvent	任职事件	shl:event	shl:Person	shl:OfficialEvent
hl:officialPosition	职衔		shl:OfficialEvent	Literal
prov:startedAtTime	开始时间		prov:Activity	Literal
prov:endedAtTime	结束时间		prov:Activity	Literal
prov:atLocation	发生地点		prov:Activity	prov:Location
hl:migrationEvent	迁徙	shl:event	shl:Person	shl:MigrationEvent
hl:originalLocality	原居地		shl:MigrationEvent	shl:Place
hl:locality	迁居地		shl:MigrationEvent	shl:Place
hl:nameType	名称类别		shl:Name	Literal
el:childOf	父母		shl:Person	shl:Person
el:parentOf	子女		shl:Person	shl:Person
el:spouseOf	配偶		shl:Person	shl:Person
el:friendOf	朋友		shl:Person	shl:Person
hl:source	来源		shl:Resource	shl:Resource
c:source	来源		shl:Resource	Literal
c:references	参考		shl:Resource	Literal
hl:relationSubject	关系主体		shl:Relationship	shl:Person
hl:relationObject	关系人		shl:Relationship	shl:Person
hl:relationType	关系类型		shl:Relationship	Literal

OpenRefine 数据混搭

- RDF->adding a reconciliation service

The screenshot shows the OpenRefine interface with a table of 1134 rows. The table has columns for id, name, gnid, county, county_gnid, city, city_gnid, province, province_gnid, town, town_gnid, village, lon, and lat. A dialog box titled "Add SPARQL-based reconciliation service" is open, allowing the user to configure a new service. The dialog includes fields for Name, Endpoint URL, Graph URI, and Type, along with a section for Label properties.

All	id	name	gnid	county	county_gnid	city	city_gnid	province	province_gnid	town	town_gnid	village	lon	lat
☆	1.	1	河南	1808520										
☆	2.	2	固始	1809216	固始县									
☆	3.	3	湖南	1806691										
☆	4.	4	株洲											
☆	5.	5	浙江	1784764										
☆	6.	6	淳安	1814787	淳安县									
☆	7.	7	遂安											
☆	8.	8	开化	1804871	开化县									
☆	9.	9	湖州											
☆	10.	10	吴兴		吴兴区									

Add SPARQL-based reconciliation service

Name:
A human readable name

Endpoint details

Endpoint URL:

Graph URI:
Leave empty to use the default graph

Type:
This determines the syntax that will be used for search

Label properties

Select properties that are used to label resources in the endpoint. These properties will be used to match resources:

rdfs:label skos:prefLabel dcterms:title dc:title
 foaf:name
 Other...

OK Cancel

OpenRefine 数据关联

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://127.0.0.1:3333/

RDF Node

Use content from cell...

The cell's content is used ...

(row index)

id

name

gnid

county

county_gnid

city

city_gnid

province

province_gnid

town

town_gnid

village

lon

lat

OK Cancel

as a URI

as text

Preview URI values

Expression

Language

Google Refine Expression Language (GREL) ▼

"http://www.geonames.org/"+substring(toString(value), 0, 7)

No syntax error.

Preview

History

Starred

Help

value	"http://www.geonames.org/"+substring(toString(value),0,7)	resolved against the base URI
1808520	http://www.geonames.org/1808520	http://www.geonames.org/1808520
1809216	http://www.geonames.org/1809216	http://www.geonames.org/1809216
1806691	http://www.geonames.org/1806691	http://www.geonames.org/1806691
null	null	null
1784764	http://www.geonames.org/1784764	http://www.geonames.org/1784764
1814787	http://www.geonames.org/1814787	http://www.geonames.org/1814787

Add another root node

OpenRefine 实体提取

1134 rows Extensions: Named-entity recognition ▾ Freebase ▾ RDF ▾

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

All	id	name	gnid	county	county_gnid	city	city_gnid	province	province_gnid	town	town_gnid	village	lon	lat
☆ ↻	1.	1	河南	1808520										
☆ ↻	2.	2	固始	1809216	固始县									
☆ ↻	3.	3	湖南	1806691										
☆ ↻	4.	4	株洲											
☆ ↻	5.	5	浙江	1784764										
☆ ↻	6.	6	淳安	1814787	淳安县									
☆ ↻	7.	7	遂安											
☆ ↻	8.	8	开化	1804871	开化县									
☆ ↻	9.	9	湖州											
☆ ↻	10.	10	吴兴		吴兴区									

Configuration — Named-Entity Recognition Extension

Services

AlchemyAPI configuration instructions

API key

DBpedia Spotlight

Confidence

Support

Zemanta configuration instructions

API key

dataTXT configuration instructions

App ID

App key

Language

Confidence

Entity type

谢谢！

Q&A

Contact: cjxia@libnet.sh.cn