Linked Open Data Instance Level Analysis Procedure

2022-01-28
Go Sugimoto

This document describes the procedure the author has executed for his Linked Open Data instance level analysis with Excel and R. Although this is experimental, it aims to provide the transparency and reproducibility of his research. By following this document, in theory, it should be possible to reproduce the previous analysis and/or modify/update the analysis with the latest information on the web (e.g. when DBpedia data has updated, a new analysis is needed accordingly).

1. Datasets

The outcomes of the analysis are provided as two versions in compressed files (Zip and 7z contain the same data) for the sake of better compatibility.

1. Excel 2006 datasets are compressed in “Excel_Datasets_Linked_Open_Data_Instance_Level_Analysis_for_Cultural_Heritage_ver2”
2. R 4.03 files are compressed in “R_Datasets_Linked_Open_Data_Instance_Level_Analysis_for_Cultural_Heritage_ver2”

The compressed file 1 (Excel) contains the folders (bold) and files as follows. The files in the Excel Data folders are omitted due to the large volume.

Excel_Macros (this folder contains all VBA files (.bas) that are explained below in this document)

- AggregateSheets.bas
- AggregateSheets_Stats_Persons.bas
- AggregateSheets_VU_CitrixEnvironment.bas
- CopyExampleSheet2AllBooks.bas
- CopyExampleSheet2AllBooks_dates.bas
- CopyExampleSheet2AllBooks_dates2.bas
- CopyExampleSheet2AllBooks_dates3_VU_CitrixEnvironment.bas
- CopyExampleSheet2AllBooks_events2.bas
- CopyExampleSheet2AllBooks_events2_xlsx3.bas

CopyExampleSheet2AllBooks_events2_xlsx3and4_mixed.bas
CopyExampleSheet2AllBooks_events5_VU_CitrixEnvironment.bas
CopyExampleSheet2AllBooks_objects.bas
CopyExampleSheet2AllBooks_objects2_VU_CitrixEnvironment.bas
CopyExampleSheet2AllBooks_persons.bas
CopyExampleSheet2AllBooks_persons2_VU_CitrixEnvironment.bas
CopyExampleSheet2AllBooks_places.bas
CopyExampleSheet2AllBooks_places2.bas
CopyExampleSheet2AllBooks_places3.bas
CopyExampleSheet2AllBooks_places4_VU_CitrixEnvironment.bas
CountVariousSheets.bas
CreateAllReportsBySources.bas
CreateMatrix.bas
CreateStatsfor4typelinks.bas
CreateStatsfor4typelinks_test.bas
CreateStatsfor4typelinks_test2.bas
CreateStatsfor4typelinks_test3.bas
CreateStatsfor4typelinks_test4.bas
CreateStatsfor4typelinks_test4_extra.bas
CreateStatsfor4typelinks_test4_extra2.bas
CreateStatsfor4typelinks_test4_extra3.bas
CreateStatsfor4typelinks_test4_extra3_allsheets.bas
CreateStatsfor4typelinks_test4_extra3_allsheets2.bas
CreateStatsfor4typelinks_test4_extra3_allsheets3.bas
CreateStatsfor4typelinks_test4_extra3_allsheets4_final.bas
CreateStatsfor4typelinks_test4_extra3_allsheets4_final2.bas
CreateStatsfor4typelinks_test4_extra4.bas
LabelCopy.bas
LabelCounter.bas
LinkCounter.bas
ListAllSheetNamesInNewSheet.bas
Module1.bas
Outgoing.bas
SheetColor.bas
TestHttp.bas
URLaggregate.bas

Excel_Example_Templates (this folder contains template sheets for data analysis)

ExampleSheet3.xlsx
ExampleSheet3_additional.xlsx
ExampleSheet4_additional_WithLinkTypes.xlsx

Excel_Data (this folder contains all the Excel 2006 data generated and used for

the analysis. Some contain graphics used in academic paper)

SourceOnly4
Dates4
Persons4
Events4
Places4
Objects4
AllReportsBySource4_3.xlsx
AllReportsBySources4_2.xlsx
AllReportsBySources4.xlsx
AllReportsBySources3.xlsx
AllReportsDates3.xlsx
AllReportsPersons3.xlsx
AllReportsPlaces3.xlsx
AllReportsEvents3.xlsx
AllReportsObjects3.xlsx
AllReportsPersons3_Outgoing.xlsx
AllReportsObjects3_Outgoing.xlsx
AllReportsPlaces3_Outgoing.xlsx
AllReportsEvents3_Outgoing.xlsx
AllReportsDates3_Outgoing.xlsx
OverallMatrix2020.csv
Dates4Matrix_3.xlsx
Dates4Matrix_2.xlsx
Dates4Matrix.xlsx
Dates3Matrix.xlsx
Persons4Matrix_3.xlsx
Persons4Matrix_2.xlsx
Persons4Matrix.xlsx
Persons3Matrix.xlsx
Events4Matrix_3.xlsx
Events4Matrix_2.xlsx
Events4Matrix.xlsx
Events3Matrix.xlsx
Places4Matrix_3.xlsx
Places4Matrix_2.xlsx
Places4Matrix.xlsx
Places3Matrix.xlsx
Objects4Matrix_3.xlsx
Objects4Matrix_2.xlsx
Objects4Matrix.xlsx
Objects3Matrix.xlsx

ListOfAllURLs.xlsx

The compressed file 2 (R) contains the folders (bold) and files as follows.

Data (this folder contains datasets generated by Excel and used for the R scripts)
R_script (this folder contains R scripts used for the analysis)
- chord.R
- networkanalysis_AllReportsBySources4.R
- networkanalysis_AllReportsBySources3.R
- networkanalysis_AllReportsBySources3\ (2).R
- networkanalysis_AllReportsBySources3_flatten.R
- networkanalysis_AllReportsBySources3_flatten\ (2).R
- networkanalysis_All_fulledges.R
- networkanalysis_Dates3Matrix.R
- networkanalysis_Events3Matrix.R
- networkanalysis_Objects3Matrix.R
- networkanalysis_Persons3Matrix.R
- networkanalysis_Places3Matrix.R

Graphics (this folder contains generated graphics. The files in the subfolders are omitted)
- **20200604**
- **20201220**
- **20211203**

Results (this folder contains Excel files holding network analysis results)
- **20210113**
  - AllReportsBySources3_stats.xlsx
  - AllReportsBySources3_stats_graphics.xlsx

2. Preparation

Note: The documentation of inverse properties are not yet automatic (this issue will be understood later in this document). It is done by manually checking ByEntitiy files (e.g. DBpedia quality analysis4.xlsx). It is possible to automate it by creating a VBA, but it is not done yet due to time constraint. It only affects DBpedia, Wikidata, and Europeana.

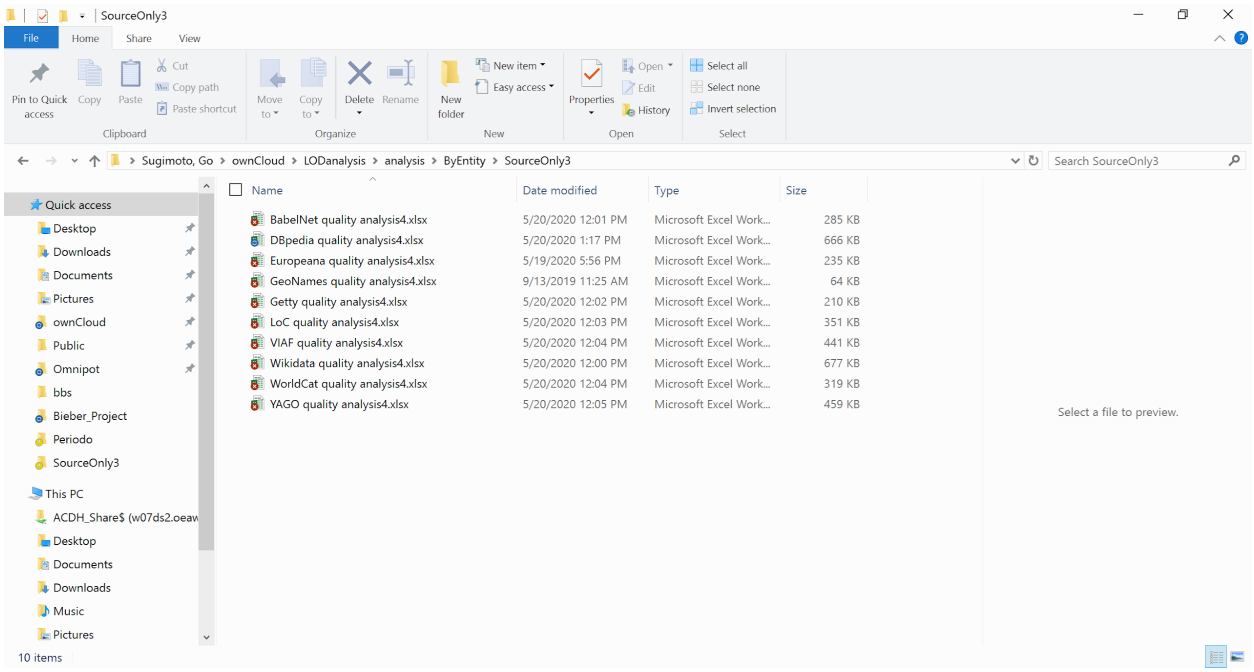
1.1 Entity identification and bookmarking

1. If needed to add new entities for the analysis, start from DBpedia to bookmark its entity URLs in the corresponding tag (DBpedia2 tag in the Firefox bookmark), because you need to check it many times later.
2. DBpedia has owl:sameAs links, which is handy to identify the entities in other data sources (e.g. YAGO, Wikidata etc)
3. If the entities are not easily found, search in the search engine and/or API of the LOD websites, because they can be hidden.
4. Bookmark two for each entity (RDF and HTML), check how they are bookmarked to make sure all are consistently bookmarked. Some entities have multi entries (or similar entities such as broader/narrower concepts and FRBR Work-Expression differences). Data quality has to be consistent across all sources, so make a decision for criteria of what to include and what not. Be careful if data looks not very normal (e.g. YAGO, VIAF are the typical usual suspects for multi coreferences)
5. Repeat the process for all dataset sources

1.2 Use of RDF Beams[1] and info filling in Excel

6. In order to save time, from the second time (e.g. to reproduce the analysis, or modify/update the analysis), it is wise to use the last version of the analysis. The old version of Excel files (Excel data structure such as cells and sheets) are available in **Excel_Data** and can be used as templates for a new analysis (i.e. do not forget to always keep backup in any case!) from C:\Users\gsugimoto\ownCloud\LODanalysis\analysis\ByEntity\SourceOnly3

[1] RDB Beams is a simple PHP web application developed to analyse lookup entities through URIs. It is not needed if you count lookup data manually, and/or with other analytical tools.



7. Keep the name and add number at the end (e.g. DBpedia quality analysis4.xlsx to DBpedia quality analysis5.xlsx) in order to distinguish the new one from the older version
8. Start from DBpedia quality analysis4.xlsx, because it has the most intuitive URLs and standardised entities (all entities available)
9. Use the latest RDF Beams http://localhost/rdfbeams/nobert_f/index_test.php (i.e. C:\xampp\htdocs\rdfbeams\nobert_f). This includes, standard RDF Beams results, plus, INVERSE properties and REGEX comparison (so that we can double-check which one is correct)
10. Access bookmarked URLs for DBpedia. View source and copy & paste RDF/XML into RDF Beams
11. Check the results (if there is a doubt, double-check with editor like Sublime or Atom with search and/or REGEX search.
12. Fill info in the tabs (it is good to select and copy the multiple-tabs, if new tabs are needed)
13. Repeat the process for other sources (Wikidata is good to do 2nd for checking)

2. Creating statistics tab for each source-based sheet (VIAF, Europeana etc)

Data

The compressed file 1 (Excel) contains VBA scripts in **Excel_Macros** and Excel data used/generated in this process.

2.1 Using Excel Macro to aggregate data

1. Enable Developer mode in EXCEL (File>Options>Customise Ribbon: Radiobox check for Developer)
2. For the preparation of Macro files (Developer > Visual Basic)

2.2 In Visual Basic for Applications Window

1. Right click VBA Project of the file > select Insert Module
2. Right click Modules folder > Import .bas File (select LabelCopy, Linkcounter, Labelcounter, SheetColor, CountVariousSheets)

2.3 Excel (normally all automatic, but manual checking and modifying may be needed)

1. Select the fist sheet (an instance e.g. Shakespeare)
2. Developer > Macros > select LabelCopy > Run (check if data is filled properly)
3. Developer > Macros > select Linkcounter > Run (check if data is filled properly)
4. Developer > Macros > select Labelcounter > Run (check if data is filled properly)
5. Developer > Macros > select SheetColor > Run (check if data is filled properly)
6. Open Wikidata quality analysis2.xlsm
7. Copy D11:F32
8. Paste it on the first sheet of your file
9. Right click the sheet tab and select all sheets
10. Home> Editing > Fill> Across Worksheets > All (check if data is filled properly)
11. Repeat 7-10, using F1:G8
12. Open Wikidata quality analysis2.xlsm
13. Go to Stats sheet. Right click the sheet tab and select Move or Copy
14. Check the create copy radiobox, select your file
15. On the newly created Stats sheet
16. Home >Find and Select>Replace [Wikidata quality analysis2.xlsm]Shakespeare:EFLCup with the range of the whole worksheets in your file (e.g. Shakespeare:EFLCup)
17. 54 cases should be replaced (check if data is filled properly)
18. Developer > Macros > select CountVariousSheets > Run (check if data is filled properly)

3. Preparation for Advanced Data Analysis with VBA

Prerequisite
All source-based sheets should be ready (including stats tab)

Aim

Create new datasets/structures to analyse the existing datasets from different perspectives. The first datasets were organised by data sources (ie DBpedia, YAGO). So, the new datasets can be organised by each entity (ie Shakespeare, London) and/or category (ie all places and persons). It is also interesting to analyse outgoing links per entity per category.

Data
The compressed file 1 (Excel) contains VBA scripts in **Excel_Macros** and Excel data used/generated in this process.

Caution
Filenames/tab names must be not too long max 30 something characters (as it returns error in Macro). Current file names are close to the limit (Europeana is named as European quality analysis4.xlsx, without “a” at the end)

Preparation
1. Create a SourceOnly file folder (in order to preserve original and use this as a working folder)

3.1 Create files for entities

1. Open a new empty Excel book
2. Open AggregateSheets.bas (See above how to deal with .bas file)
3. Modify this part for your needs

```
#Here specify the folder to create
Const SOURCE_DIR As String =
"C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\"

#Here specify the new file (ie CLinneaeus.xlsx is the file to
aggregate data about Linneaeus from all the files in the folder above
(SOURCE_DIR), e.g. BabelNet quality analysis2.xlsx)
Const DEST_FILE As String =
"C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\Persons\CLinn
eaeus.xlsx"
#Here specify which tab should be aggregated from the files in the
SOURCE_DIR e.g. BabelNet quality analysis2.xlsx. In this case "C
Linneaeus" is the tab name to be collected)
sWB.Worksheets("C Linneaeus").Copy After:=dWB.Worksheets(dSheetCount)
```

4. IMPORTANT: Before starting, check if all entities are available for all files in SOURCE_DIR (e.g. Persons are not available in GeoNames quality analysis2.xlsx, so temporally remove this file to generate the person entity files)

5. Run AggregateSheets.bas. CLinneaeus.xlsx should be created with tabs from different source files (file names are used for tab names)
6. Repeat the process for all entities (manually)
7. When all entities are created for a category, repeat the process for another category

3.2 Organise the entity files by batch processing (add empty tabs, Aggregate and Stats tabs for consistency purpose)

When all categories are done, we will create an additional aggregation dataset (i.e. “Aggregate” and “Stats” tabs in the beginning of a file) for each entity, by calculating the sum of each tab of it. To facilitate this, the order to tabs is fixed. It should be **YAGO-WorldCat-Wikidata-VIAF-LoC-Getty-GeoNames-European-DBpedia-BabelNet-Wikipedia**. If this order breaks (especially start and end), automation will be disturbed. Stick with it. In case of additional tabs (such as VIAF2 for Australia), consolidate info by copy and pasting to the primary entity *before this macro*, so recalculation is avoided. To do this, go to category files (e.g. VIAF quality analysis2.xlsx) and manually add stats for the following duplicate entities for consolidation

- LoC >> King2, Tamil2, Tamil3
- VIAF >> Alex2, Aristotel2, Australia2, J Caesar2, Pitti2, Sgt2, Byzantine2, King2, King3
- ~~WorldCat >> WWII PERIOD and WWI PERIOD~~ (Omitted from the consolidation, because they are not exactly the duplicate, and different entities)
- YAGO >> India2, Spain2, London2, Russia2, California2, NewYork2, Netherlands2, Sweden2, Olympic2, Tamil2

In case of missing tabs, ignore for the time being. They should be taken into account later.

8. Open CopyExampleSheet2AllBooks.bas (or more specifically customised VBA: CopyExampleSheet2AllBooks_Dates.bas)
9. Modify this part for your needs

```
#Here specifies the source folder, which includes entity files such
as CLinneaeus.xlsx
Const SOURCE_DIR As String =
"C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\Test\SourceOn
ly\Persons\"

#Here specify the source template file, from which we will copy some
tabs to the destination files (No need to change, unless rework is
done in the previous sections). Current vesion is
C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\ExampleSheet3.
xlsx
```

```
Const SOURCE_FILE As String =
"C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\ExampleSheet3
.xlsx"

#Here specifies missing tabs (i.e. the sources which do not have any
entities tend to be omitted from the files in the source directory).
To make all the files consistent, we create tabs in which data have
all 0 values. This makes our tasks more automatic later. To this end,
first, examine some files in the category you work on, and check what
tabs are missing. Then, add or remove tabs in this VBA. Wikipedia tab
is most likely to be added. Note Wikipedia and Europeana have too
long names “sometimes”, so it should be “xls” not “xlsx”
(unfortunately Excel sheets are not very consistent and I gave up
making it perfect). Do double-check the outcome. Some source-entity
files (e.g. Getty) may have empty tabs already, which means, this
macro may duplicate the tabs, which is not needed. You can use syntax
such as “Before” and “After” as well as fine names in the last
parenthesis (). to specify the order of the newly creating tab. Don’t
forget which files should be in the folder, because unneeded file
stops the script, which may freeze Excel and restart of Windows may
be required.

dWB.Worksheets("Wikipedia quality analysis4.xls").Copy
After:=sWB.Worksheets(dSheetCount)
dWB.Worksheets("European quality analysis4.xls").Copy
Before:=sWB.Worksheets("DBpedia quality analysis4.xlsx")
dWB.Worksheets("GeoNames quality analysis4.xlsx").Copy
Before:=sWB.Worksheets("European quality analysis4.xls")
dWB.Worksheets("Getty quality analysis4.xlsx").Copy
After:=sWB.Worksheets("LoC quality analysis4.xlsx")
```

10. Run CopyExampleSheet2AllBooks.bas (or more specifically customised VBA: CopyExampleSheet2AllBooks_Dates.bas) For example, new tabs should be added in CLinneaeus.xlsx. **CopyExampleSheet2AllBooks.bas will automatically modify all files in a category.** They should at least include “Aggregate” and “Stats” tabs at the beginning. A few missing tabs specified above will be added at the end of the tabs.
11. Repeat this process for each category (Dates, Events, Places, Persons)
12. Open each xlsx file (eg CLinneaeus.xlsx) and go to “Aggregate”
13. Home >Find and Select>Replace C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\[ExampleSheet.xlsx] with empty value. So, the syntax should look like 'YAGO quality analysis2.xlsx:Wikipedia quality analysis2.xls'!D12)

14. 200 cases (220 cases in case of including Europeana)should be replaced (check if data is filled properly)
    When there is an error, it may be because Europeana tab name is different (xlsx or xls). Just manually modify the tab name to xls. List of problematic entities are:
    WWI, Aristotle, Beethoven,Hitler, Mozart, Prato, Shakespeare, Madonna, Australia, Canada, England, France, Germany, India, Italy, Japan, Netherlands, NewYork.
15. Go to “Stats” tab and repeat 12-13. 64 cases should be replaced (check if data is filled properly)
16. Save the change
17. Repeat the 11-15 for each entity

3.3a Create AllReports.xlsx by aggregating data from all entities

Let’s reuse AggregateSheets.bas to create AllReports.xlsx. Paths to the source folder and target file name (e.g. in the upper folder from the source folder) have to be changed, and tab name (“Stats”), but all other parts are the same.

18. Open AggregateSheets.bas (or more specifically customised VBA: AggregateSheets_Stats_Persons.bas)
19. Modify this part for your needs

```
#Here specifies the folder of a category (ie Persons)
Const SOURCE_DIR As String =
"C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\Test\SourceOn
ly\Persons"
#Here specifies the output file in an upper folder (conventionally
always called AllReports2.xlsx)
Const DEST_FILE As String =
"C:\Users\gsugimoto\ownCloud\RDFBeams\analysis\ByEntity\Test\Persons\
AllReports.xlsx"
# Here specifies the tab to copy (ie. always “Stats”)
sWB.Worksheets("Stats").Copy After:=dWB.Worksheets(dSheetCount)
```

20. Run AggregateSheets.bas (or more specifically customised VBA: AggregateSheets_Stats_Persons.bas). This should generate AllReports2.xlsx, the summary of all entities in the category
21. Open the latest version of AllReports.xlsx (e.g. AllReportsDates3.xlsx) from another category and the new version of AllReports.xlsx just created
22. Go to “Stats” and “Aggregate” sheets in AllReports.xlsx. Right click the sheet tab and select Move or Copy to the left-most position of all the tabs.
23. Check the create copy radiobox, select your file (ie AllReports2.xlsx)
24. Go to AllReports2.xlsx
25. Examine the start and end tabs

26. Home >Find and Select>Replace [AllReport.xlsx]1968:1980] with the range of the whole worksheets in your file (e.g. Shakespeare:AlexGreat)(64 or 66 (with Europeana) changes for “Stats”
27. (Optional for Europeana case: Manually manipulate the formulas in the newly created “Aggregate” tab, based on the Aggregate tab of the latest version of AllReport.xslx (e.g. AllReportsDates3.xlsx)
28. Save as AllReports.xlsx (Unfortunately the same name files cannot be opened at the same time (to do 21). This is why this workaround is needed to rename the file)
29. Repeat the whole process (2-) for each category

3.3b Generate Aggregate and Stats tab for 4 link types for each category

All categories can be produced as described above 3.3 (18-28), but procedure for each category is different. Do not follow Section 3.3a (18-29).

- Create a new folder and put all source files (e.g. BabelNet quality analysis5.xlsx).
- Open each (e.g. BabelNet quality analysis5.xlsx) and import ListAllSheetNamesInNewSheet.bas use. First, run Sub GetSheetList() in the bas file. It will create a list of all sheets within. Second, manually edit the order of list in the A column (or reuse prepared list in full in ListSheetNames.xlsx). Third, run Sub ChangeOrder() to sort the order of the sheets.
- Go to Stats sheet to calculate the sum of the sheets for each category, which is now in the sequence within the file. Copy template tables from Stats sheet (U to Y column). It is needed to change the range of the sheets of a category (e.g. Shakespare:QueenVictoria). This will generate the right tables for all link types (e.g. skos:exactMatch) for each category (e.g. BabelPerson) in the source file (e.g. BabelNet quality analysis5.xlsx)
- Save the file as AllReportsPersons4.xlsx etc
- Repeat the previous process for each source (e.g. YAGO quality analysis5.xlsx)

3.4 Create Matrix.xlsx for each category by aggregating Aggregate tabs from each entity

30. Simply open CreateMatrix.bas
31. Modify this part for your needs

```
'Specify below the folder in which all input files are present'
Const SOURCE_DIR As String =
"C:\Users\gsugimoto\ownCloud\LODanalysis\analysis\ByEntity\Objects3\"
```

```
'Specify below the output file
Const DEST_FILE As String =
"C:\Users\gsugimoto\ownCloud\LODanalysis\analysis\ByEntity\Objects3Ma
trix.xlsx"
```

32. Reuse the existing aggregation tab for this purpose (e.g. Dates3Matrix.xlsx)
33. Copy the Aggregation tab to the new Matrix file (e.g. Dates4Matrix.xlsx)
34. Examine the start and end tabs of the new Matrix file
35. Go to Aggregate tab
36. Home >Find and Select>Replace [Dates3Matrix.xlsx]1987.xlsx:1960.xlsx with the range of the whole worksheets in your file (e.g. 1987.xlsx:1960.xlsx)242 cells will be changed
37. Repeat the process for each category (e.g. change No 31 for Persons)

3.5 Create AllReportsBySources.xlsx to have the overview of all stats by sources

38. Simply open AllReportsBySources.bas
39. Modify this part for your needs

```
'Specify below the folder in which all input files are present'
Const SOURCE_DIR As String =
"C:\Users\gsugimoto\ownCloud\LODanalysis\analysis\ByEntity\SourceOnly
3\"
     'Specify below the output file
Const DEST_FILE As String =
"C:\Users\gsugimoto\ownCloud\LODanalysis\analysis\ByEntity\AllReports
BySources3.xlsx"
```

Make sure below is “Stats”

```
sWB.Worksheets("Stats").Copy After:=dWB.Worksheets(dSheetCount)
```

40. Reuse the existing aggregation tab for this purpose. Copy the Aggregation tab to the existing AllReportsBySources3.xlsx file
41. Examine the start and end tabs of the new AllReportsBySource file
42. Go to Aggregate tab
43. Home >Find and Select>Replace old links with the range of the whole worksheets in your file (e.g. YAGO quality analysis4.xlsx:Wikipedia quality analysis4.xls).

3.6 Create outgoing links summary

In the previous sections, the focus of the analysis is link destinations (i.e. to which source links are going). This section aims to count the amount of links found in each entity/source/category

(i.e. how many outgoing links are held in each data source). The outgoing links include links to any sources including 11 sources.

1. Create a copy of category folder(to make sure original data is preserved) and apply VBA for the files in this new folder)
2. Apply Outgoing.bas for each entity (ie file) in each category, which will add “outgoing” tab at the beginning of each file
3. Aggregate 1) to generate a summary table for each category
4. Generate charts from the table for analyses

Note: 3.6 is not needed, as it will generate a large amount of Matrix tables (e.g. 4 linktype tables for each category i.e. 4*5 tables in total)

4. Using the Excel outcome files for R analysis

Aim

After generating EXCEL files, important stats will be used for R network analyses. For this, the overall matrix will be used, as well as Matrix files created above for each category

Data

The compressed file 2 (R) contains the R scripts in **R_script** and the source data used in **Data 20201220**. The results are available in **Graphics** and **Results.**

Preparation

This section is obsolete. The latest R scripts can read XLSX files directly and CSV is not needed.

1. ~~The overall matrix is already present in the aggregate tab of AllReportsBySources.xlsx. To make it clean, just check the duplicate row/column (ie YAGO) and make sure the matrix table is consistent~~
2. ~~To avoid modifying the original AllReportsBySources.xlsx, create a separate CSV file, only consisting of matrix data from 1~~
3. ~~Import CSV into R~~
4. ~~Similarly, repeat the same process for category matrix files (e.g. Dates2Matrix.xlsx)(i.e. apply 2 and 3 above).~~

Script execution

You may need to install or update libraries listed on the top of each R script. When executing scripts, you may see errors. You may need to double-check R and used library versions, because R scripts may not work because of problems with dependencies.

1. To generate chord diagrams, simply open and run Chord.R
2. To generate network diagrams and generate network analysis statistics, simply

open and run networkanalysis_xxx.R files

Appendix1. Create stats for each type of links (e.g. owl:sameAs and rdfs:seeAlso)

1.Preparation

Filling information to distinguish four types of links for analysis (skos:exactMatch, rdfs:seeAlso, owl:sameAs, schema:sameAs) Positions of cells are important later for aggregated calculations, so do not change!

1) Excel on VU Citrix. Use the uploaded EXCEL files on OneDrive
2) Open each entity file (e.g. AlexGreat.xslx)
3) Go to 1st tab (i.e. YAGO quality analysis4.xslx tab)
4) Copy D10 to D36 and paste at U10 with functions.
5) Check Column B and manually fill data from U10 to Y36
6) Repeat 2) to 4) for all tabs (i.e. until Wikipedia qualit analysis4.xslx tab)
7) Go to Stats tab and select U10 cell. Apply SUM=(), then click the start tab (i.e. YAGO quality analysis4.xslx tab) and select U10. Then click the end tab (i.e. Wikipedia qualit analysis4.xslx tab) with shift pressed. This will select all tabs from YAGO to Wikipedia. Press Enter. This will calculate SUM for U10 of all tabs for U10 in the Stats tab. Select U10 in the Stats tab and drag to Y36. This will extend U10 function to all cells until Y36, meaning SUM are calculated from U10 to Y36.
8) Go to Aggregate tab. Select and copy A2 to N26 (all cells in 2 tables). Paste it at R2. Put name in R1 (e.g. this table is for skos:exactMatch). Select T4 to AE26 (all cells with numeric values). Replace D with V (220 will be replaced). Do not use Find first, because it will deselect the selected cells. This makes calculation for skos:exactMatch. Check if the numbers are right
9) Repeat 7 for rdfs:seeAlso (start from AG2), owl:sameAs (start from AV2), schema:sameAs (start from BK2)
10) **Save the file** (Unlike cloud Excel in Microsoft or Google, Excel on the VU’s Citrix Virtual Workstation will not automatically save the file)
11) **Let’s aggregate each entity (follow the same procedure from previous Sections)**

2. Create stats for each type of links

<Datasets to be used>
SourceOnly3 folder contains the latest aggregated data for each entity, reflecting Outgoing links (so, no need to do from earlier stage of files). Note that, although the last modified dates are newer, SourceOnly3_backupBeforeFinalising folder has older numbers. So, let's use files from SourceOnly3 folder an**d** generate the following ones with macros etc.

For overall stats

- **AllReportsBySources3.xlsx** (does not include INVERSE links. This is to be used for R visualisation)
- **AllReportsBySources3 - copy.xlsx** (includes INVERSE links. This is also to be used for R visualisation for comparison)

For stats in each category

- **AllReportsPersons3 _outgoing.xlsx** (includes outgoing links, so it is the one to use for each category for R visualisation)
- **AllReportsPerson3 -Copy.xlsx** (does not include outgoing links)