WP5: Tool tests for scenario 1.3

Authors:

Eva Toller (National Archives of Sweden, RA)
……

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	**Public**	**P**
C	**Confidential, only for members of the consortium and the Commission Services**	

Revision History

Revision	Date	Author	Organisation	Description
0.1	20130503	Eva Toller	RA	First draft
0.2	20130506	Eva Toller	RA	Added tests results for ROND (Riksarkivet Open Data).
0.3	20130618	Eva Toller	RA	Added “Quality of result” as another quanification factor

1 TOOL TESTING FOR SCENARIO 1.3

1.1 SCENARIO DESCRIPTION

“A little museum in Malta has a historical library and a digitised personal archive collection. The museum has staff of only 9 and only voluntary IT support. The director of the museum is aware of the need to organise digital preservation for the digitised documents, but is not sure how to do it. He receives periodically offers for long-term storage of digital content, but finds it difficult to select or to make a decision. He has practically no IT competence to rely on for decision-making, but is convinced that the decision should be forward-looking and accommodate the needs of the museum for the next 5 years.”

General comment: if this scenario is reused in Proof of Concept #2, we could try to find a *real* organisation that has this problem (although it does not have to be a museum).

Suggested test data: see document **DCH-RP_WP5_Scen-1-3_ID-51.pdf**

1.2 DISPOSITION

Chapter 2 and the following chapters are structured in the following way:

In sections X.1, a short description is given of the tool and how it works.

In sections X.2, the data set(s) that the tool will be tested on is described. If there are several data sets, they are described in sub sections: X.2.1, X.2.2, X.2.3 … X.2.n.

In sections X.3, the results of the tests are given. If there are several data sets and the results differ significantly between them, they are described in sub-sections: X.3.1, X.3.2, X.3.3 … X.3.n.

In sections X.4, general comments are given about the tool and its usability for digital cultural heritage preservation, dissemination et c.

2 ROND

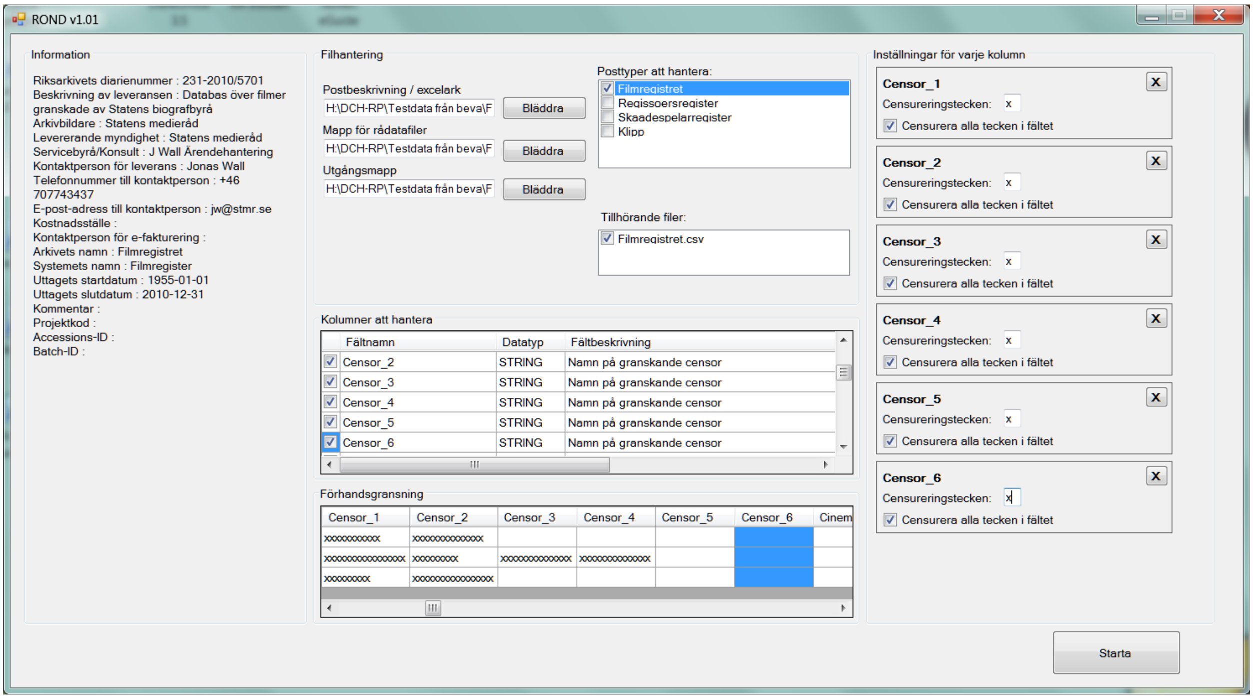
2.1 GENERAL DESCRIPTION

ROND (Riksarkivet OpeN Data) is a tool for de-identifying data sets. It is written in C#. It is dependent on Riksarkivet's chosen meta data format for structured text files, **ADDML** : http://xml.ra.se/addml/ (in Swedish). ROND was developed at Riksarkivet in 2012, using a grant from Vinnova; see http://www.vinnova.se/sv/Resultat/Projekt/Effekta/Riksarkivet-Open-data-pilot---RONDp/ (in Swedish).

ROND works in the following way. You choose a meta data file (currently, an Excel spreadsheet) and a directory where the corresponding raw data files are situated. Then you give a directory where the de-identified files should be written.

The data is loaded, and now you can choose a record type corresponding to a specific file (“Posttyper att hantera”) and choose what columns that should be de-identified (“Kolumner att hantera”). The default character for subsitution of the de-identification candidates is 0 (zero), but you can choose any other suitable character (“Inställningar för varje kolumn”, “Censureringstecken”). You can also choose to either substitute all characters in the field (“Censurera alla tecken I fältet”), or only some of them. The latter is feasible when you, for example, have social security numbers where you want to keep the birth year and birth month only.

The output files that are changed are named <old filename>**_konverterad.**<file extension>



Information

Riksarkivets diarienummer : 231-2010/5701
Beskrivning av leveransen : Databas över filmer granskade av Statens biografbyrå
Arkivbildare : Statens medieråd
Levererande myndighet : Statens medieråd
Servicebyrå/Konsult : J Wall Ärendehantering
Kontaktperson för leverans : Jonas Wall
Telefonnummer till kontaktperson : +46 707743437
E-post-adress till kontaktperson : jw@stmr.se
Kostnadsställe :
Kontaktperson för e-fakturering :
Arkivets namn : Filmregistret
Systemets namn : Filmregister
Uttagets startdatum : 1955-01-01
Uttagets slutdatum : 2010-12-31
Kommentar :
Projektkod :
Accessions-ID :
Batch-ID :

Filhantering

Postbeskrivning / excelark: H:\DCH-RP\Testdata från beva\F
Mapp för rådatafiler: H:\DCH-RP\Testdata från beva\F
Utgångsmapp: H:\DCH-RP\Testdata från beva\F

Posttyper att hantera:
☑ Filmregistret
☐ Regissoersregister
☐ Skaadespelarregister
☐ Klipp

Tillhörande filer:
☑ Filmregistret.csv

Kolumner att hantera

	Fältnamn	Datatyp	Fältbeskrivning
☑	Censor_2	STRING	Namn på granskande censor
☑	Censor_3	STRING	Namn på granskande censor
☑	Censor_4	STRING	Namn på granskande censor
☑	Censor_5	STRING	Namn på granskande censor
☑	Censor_6	STRING	Namn på granskande censor

Förhandsgransning

Censor_1	Censor_2	Censor_3	Censor_4	Censor_5	Censor_6	Cinem
xxxxxxxxxxx	xxxxxxxxxxxxxx					
xxxxxxxxxxxxxxxx	xxxxxxxxx	xxxxxxxxxxxxxx	xxxxxxxxxxxxxx			
xxxxxxxxx	xxxxxxxxxxxxxxxx					

Inställningar för varje kolumn

Censor_1
Censureringstecken: x
☑ Censurera alla tecken i fältet

Censor_2
Censureringstecken: x
☑ Censurera alla tecken i fältet

Censor_3
Censureringstecken: x
☑ Censurera alla tecken i fältet

Censor_4
Censureringstecken: x
☑ Censurera alla tecken i fältet

Censor_5
Censureringstecken: x
☑ Censurera alla tecken i fältet

Censor_6
Censureringstecken: x
☑ Censurera alla tecken i fältet

Figure 1: An example of de-identification in ROND

ROND does the following changes to the files:

1. In the meta data file (Excel spreadsheet), comments are inserted about the columns that have been de-identified. If the size of the file has been changed as a direct consequence of the de-identification, this is noted. The new names of the de-identified raw data file are written instead of the original file names.
2. Besides from de-identifying the chosen columns, ROND also inserts row numbering after the rightmost column in the changed raw data files (consecutive numbering 1,2,3, … , n). This is done to later make it possible to identify the changed data when needful, by comparing the <old filename>**_konverterad.**<file extension> file with the original file.

2.2 DATA SET

The data set that was used to test ROND is called “Filmregistret” (the Film Records Collection). These are not records of films *per se*, but of the censorship that has been performed for some films containing scenes that are violent or offensive in other ways. The actual cuts are not included in this data set.

There are four separate files in “Filmregistret”:

Filmregistret.csv: contains general and administrative information about the films and the censoring process (including the name of the censors and the technicians). All in all, there are 109 columns in this file. The number of record instances (rows) is 59785.

Regissoer.csv: contains the names of the directors and IDs for the censorships for that director's films. There are only these 2 columns in this file. The number of record instances (rows) is 30405.

Skaadespelare.csv: contains the names of the actors and IDs for the censorships for that actor's films. There are only these 2 columns in this file. The number of record instances (rows) is 139301.

Klipp.csv: contains detailed information about the cuts that have been made (description of the scenes, lengths of cuts, sections of law that are referred to, and so on). There are 14 columns in this file. The number of record instances (rows) is 8330.

2.3 TEST RESULTS

Although “Filmregistret” does not, “by the letter of the law”, contain any confidental information, it was considered prudent to de-identify the names of the censors and the technicians. The names of the directors and actors were not de-identified; firstly, they are widely known, and secondly, to remove them would greatly decrease the usability of the data set.

Names of censors and technicians are columns in *Filmregistret.csv:*

Censor_1, Censor_2, Censor_3, Censor_4, Censor_5, Censor_6

Tekniker, Teniker2

All letters in the names were replaced with x's (see Figure 1 in section 2.1 for an example).

The program behaved as expected and gave the correct results. However, there may be a problem with this program concerning usability.

2.4 USABILITY

There is no manual and no other help for the user, since it was deemed to be relatively easy to use by the developers. However, some of the text that is used in labels is quite misleading, and it is also unnecessarily cumbersome to choose the files you want to work with.

It would be rather easy to fix these problems, and then the program would be (even more) easy to use, also for a non-technician. It is also recommended that some in-built help is added, for example “tool tips” that contain short explanations about the different things you must do in order to get the de-identification running.

ROND is easy to install; it is a C# program that is compiled to an .exe file that can be easily distributed and/or downloaded.

Recommendation

If ROND should be recommended as a tool for de-identification of archive information in the DCH sector, the improvements mentioned above should be made first. It should also be pointed out that ROND has a major limitation in that it requires a certain metadata model (ADDML), which is currently only used by Sweden and Norway.

However, tools of this *type* could be very useful for publishing huge amounts of archival information as open data; information that otherwise would be locked up in the archives, and much harder to find and obtain for interested parties.

Grade

On a scale from 1 (very bad) to 5 (very good). X is “Not applicable”.

Simplicity of installation: 5

Simplicity of management: X

Ease of use: 2 - 3

Generality of solution: 1

Quality of result: 5