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 quantification factor
0.4	20130729	Eva Toller	RA	Added results for Archivist's Toolkit (which was not possible to run since it was not possible to fully install it).
0.5	20130807	Eva Toller	RA	Added results for Xena.
0.6	20130808	Eva Toller	RA	Modified some of the text for Xena after comments from the Product Owner.

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 any). 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. (This section may be skipped if it was not possible to install and/or run the tool).

1.3 TEST ENVIRONMENT

When nothing else is said, the test environment is a PC (Personal Computer) with Windows 7 Professional, processor Intel(R) 2,7 GHz, and 8 GB working memory (RAM).

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>

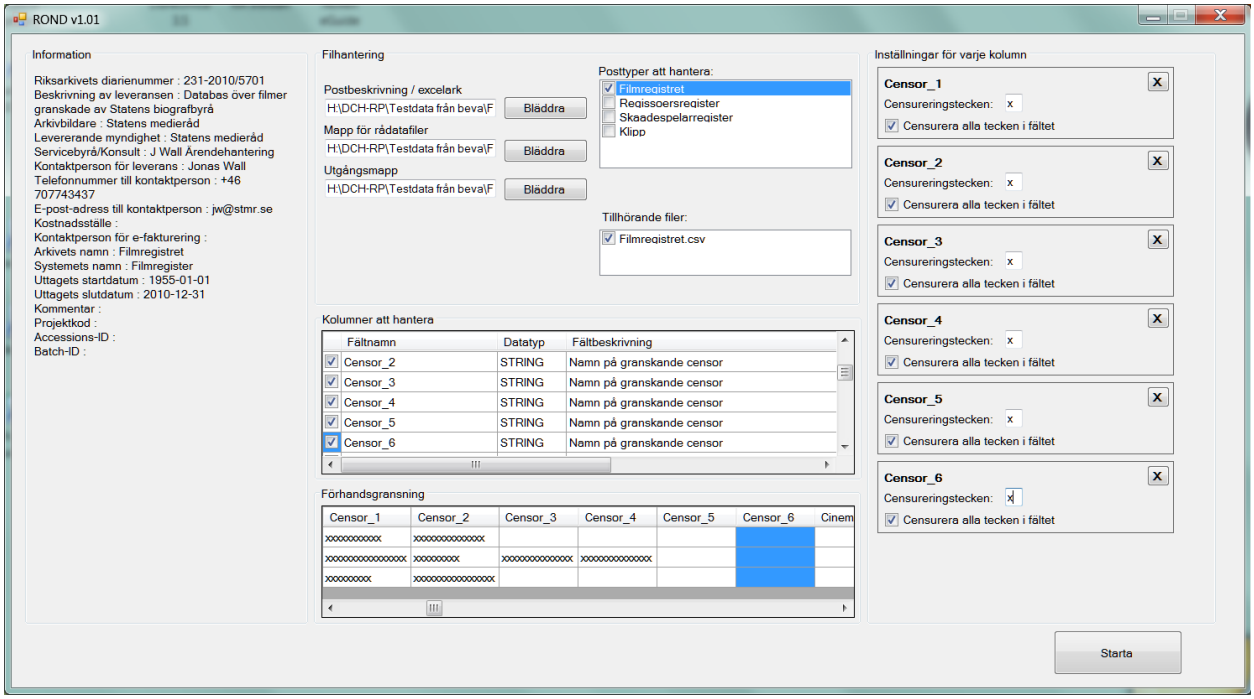

Figure 1: An example of de-identification in ROND

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.

2.4.1 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

3 ARCHIVIST'S TOOLKIT

3.1 GENERAL DESCRIPTION

Archivist's Toolkit can be downloaded from http://archiviststoolkit.org/download/release/2_0

A short description: “The Archivists’ Toolkit™, or the AT, is the first open source archival data management system to provide broad, integrated support for the management of archives. It is intended for a wide range of archival repositories. The main goals of the AT are to support archival processing and production of access instruments, promote data standardization, increase processing efficiency, and lower training costs.”

Archivist's Toolkit is available for Mac OS X, Linux, and Windows (Windows is the recommended option). It can be downloaded either with or without Java VM.

3.2 DATA SET

The data set that was to be used was the same as for ROND (see section 2.2).

3.3 TEST RESULTS

Not applicable.

3.4 USABILITY

The downloadable file is named **installArchivistsToolkit2_0u14_NoVM.exe**. As usual, you doubleclick on the downloaded file to install the program. The installation then proceeds with **InstallAnywhere**. It was recommended that you close all other programs before installation.

The usual pop-up windows that can be expected during program installations are shown:

- Acceptance of Licence Agreement
- Selection of the directory for installation
- Where to place shortcut folder

The first time you start the program, the following pop-up window is shown:

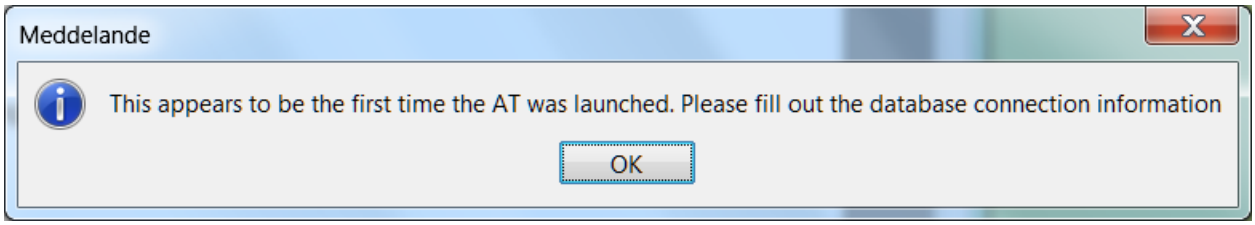

Figure 2: Pop-up window shown at program start

Now, not earlier, it is obvious that you need a database system installed: Oracle, MySQL, MS SQL Server, or “Internal Database” (however, this requirement is listed in the user manual, http://archiviststoolkit.org/sites/default/files/AT1_5_UserManual.pdf).

It is now hard to quit the program it you don't provide the information about the database (unless you restart your computer).

MySQL Community Server (MSI Installer) can be downloaded from http://dev.mysql.com/downloads/ . First, you have to create an Oracle account, and supply miscellaneous information (for example, intended use).

When trying to connect MySQL to the Toolkit, I made an error (supplied the wrong type of database). I then uninstalled Archivist's Toolkit, re-started the computer, and re-installed the Toolkit. But apparently, the Toolkit can not “forget” the previous installation error, so it is not possible to run the program at all.

Archivist's Toolkit and MySQL could be re-installed on another computer. It may also be possible to manually remove the information that causes the ToolKit to be deadlocked. However, doing this is probably beyond the scope of the museum director in Scenario 1.3.

On a scale from 1 (very bad) to 5 (very good):

Simplicity of installation: 1-2

4 XENA

4.1 GENERAL DESCRIPTION

“Xena software aids digital preservation by performing two important tasks: detecting the file formats of digital objects and converting digital objects into open formats for preservation.”

4.2 DATA SET

Samples from test data for Scenario 1.3 (see **DCH-RP_WP5_Scen-1-3_ID-51.pdf**) and from Scenario 2.2 (see **DCH-RP_WP5_Scen-2-2_ID-66-restricted.pdf**).

4.3 TEST RESULTS

For normalisation (XML as output, quality verified in the Xena Viewer):

- JPEG: very good
- TIFF: very good (except for one gigantic file that couldn't be processed within reasonable time limits)
- csv: good (one of the files couldn't be converted)
- Open Document Text: not so good (Swedish characters were omitted)
- DjVu: probably bad (the result couldn't be viewed in the Xena Viewer)
- Excel: bad (couldn't be converted although Xena recognised the format)

For binary normalisation (producing XML as output):

- Worked for all formats, but the quality of the result is not verifiable in the Xena Viewer

For conversion:

- Worked only for TIFF (which was converted to PNG)

4.4 USABILITY

4.4.1 Download and installation

Xena can be downloaded from http://xena.sourceforge.net/ . On this site, it is written about the Window version that the installation includes all necessary software except LibreOffice (which can be downloaded from http://www.libreoffice.org/).

However, from the directly reachable download page, it is unclear which version you should download to get the correct version for Windows (if any).

If you go from there to the Wiki (http://sourceforge.net/p/xena/wiki/Main_Page/), there is documentation about how to install and run Xena. Under the headline "Downloading Xena", there are finally listed different versions explicitly for Linux, Windows, and Mac OS. (Apparently, the Windows download is an **.exe** file).

The default directory for installation under Program is "**National Archives of Australia**/Xena/" (you may want to change this). Otherwise, the installation was very simple.

After the installation is finished, a README.txt file is automatically opened. It contains short descriptions of useful things to know (such as: there is a shortcut to Xena under the Start menu, how to set the output directory, and so on).

4.4.2 Converting files

The start page of Xena looks like this:

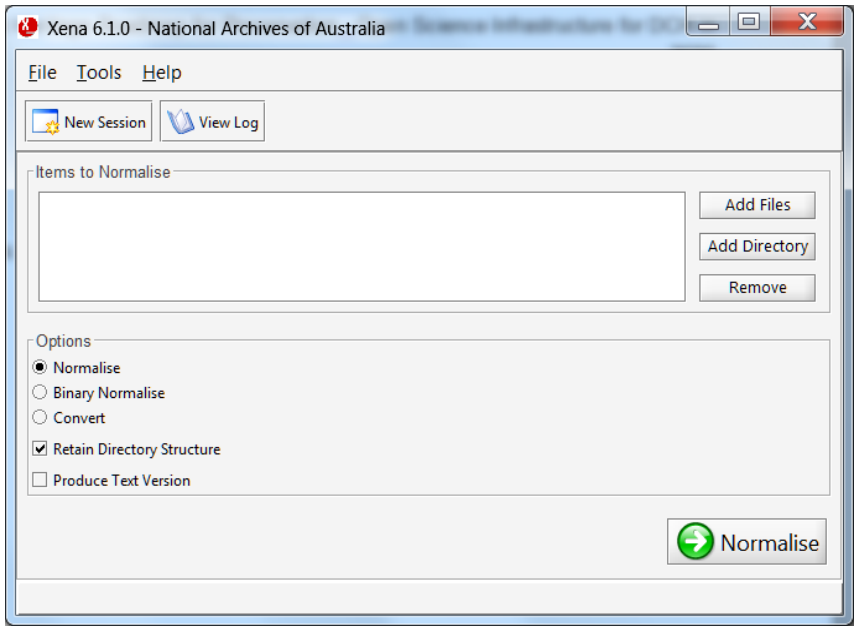

Figure 3: Start page of Xena

It is also possible to run Xena via the command line interface, if you prefer that (double-click **xena.bat**).

On http://xena.sourceforge.net/, the following is stated:

"Xena software aids digital preservation by performing two important tasks:

- detecting the file formats of digital objects
- converting digital objects into open formats for preservation."

It is not stated *which* "open formats for preservation" that are supported. However, the Help menu in Xena lists what can be done:

1. Normalise one or more files
2. Normalise a directory (that is, all files in that directory)
3. Binary normalising
4. Convert

At least when converting files, you also have the option to produce a text version.

Normalising one or more files

To choose test data, click the “Add files” button to the right.

For a first test, a medium-sized TIFF file was chosen: **NAS_Post_cards_KRA002.tif** (approximately 1,5 MB). The button “Normalise” (in the south-east corner of the window) is then clicked. The system works for some seconds, then the following message is shown:

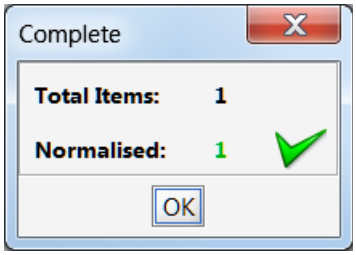

Figure 4: Normalisation was successful

The output file is (automatically) named **NAS_Post_cards_KRA002.tif_Image Tiff Normaliser.xena**. If you try to normalise the same file several times, the previous files are not overwritten, but the new output files are named **NAS_Post_cards_KRA002.tif_Image Tiff Normaliser*0001*.xena**, **NAS_Post_cards_KRA002.tif_Image Tiff Normaliser*0002*.xena**, et c.

The “.xena” format is an XML format. If you double-click on a .xena file, the Xena Viewer is opened:

Figure 5: The Xena Viewer

...and a rendering of the image is shown in another window:

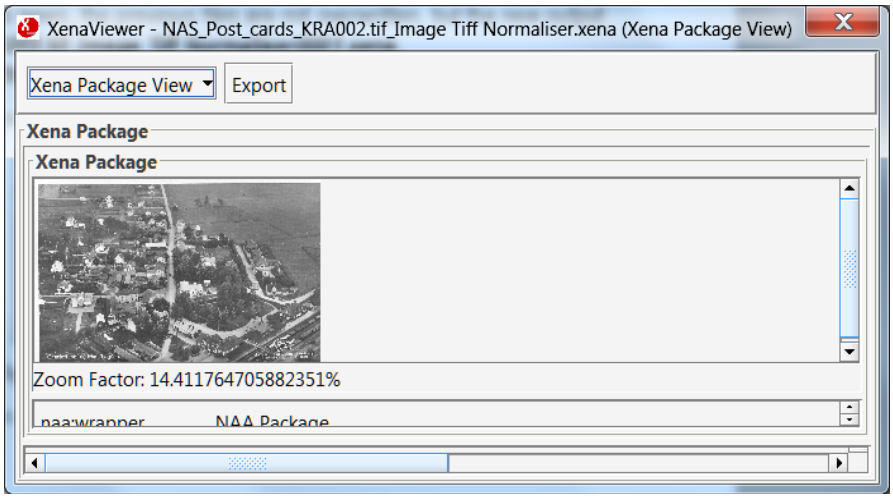

Figure 6: The normalised image

The size of the output file (1,8 MB) was slightly larger than the size of the input file (1,5 MB).

The xml code can also be viewed in the Xena Viewer (choose “Raw XML View” instead of “Xena Package View”).

I tried to normalise the largest TIFF file, **NAS_Svenska_Inskrivningskartan_image017_PS8_autocontrast-autocolor.tif** (407 MB): after fifteen minutes, it was still not finished. The next largest file, **NAS_Geometric_maps_NAS_agk_a9_090.tif** (154 MB) took less than two minutes to complete (the output file was 117 MB).

As mentioned above, one of the objectives of Xena is to automatically detect (or guess?) the file formats of digital objects. For the TIFF files, Xena correctly detected what they were. For other file types, the “guesstimates” and general normalisation results were the following:

Input file format	Xena's guessed file format	Correct guess?	Normalisation succeeded according to Xena	Normalisation succeeded for real
Tiff	Tiff	Yes	Yes (but not for the largest file???)	Yes (but not for the largest file)
DjuVu	Binary	Yes, but too general a type???	Yes	No (only seemed to do so???)
Excel (xls)	Excel	Yes	No	No
Open Document Text (odt)	ODF Document	Yes	Yes	Partly (omitted the Swedish letters **å, ä, ö**).
csv	csv	Yes	Partly (succeeded with all but one of the files)	Partly (succeeded with all but one of the files)
jpeg	jpeg	Yes	Yes	Yes

Table 1: Overall performance of Xena

When testing normalisation of several files simultaneously, the program behaved as expected (however, it didn't always succeed with all the files, as for the csv files and tiff files).

Normalising a directory

A directory that contained three jpeg images was chosen. The program behaved as expected and converted the jpeg files, but complained about the **thumbs.db** file (this is a file that is auto-generated by Windows, and normally not visible).

Then, a directory that contained seven directories, that altogether contained twelve jpeg files, was chosen. The program could handle several levels, and converted all twelve jpeg images. Again, it complained about thumbs.db. It also normalised an autogenerated and invisbile file, **desktop.ini** (from plain text to plain text).

Binary normalising

Binary normalisation preserves the original format of a file without converting it to an openly specified format. I tried to binary normalise all the 145 files at once (except the very large TIFF file). Xena doesn't try to guess the file format for binary normalisation.

Xena normalised all 144 files without problems. These files are not meant to be viewed in the Xena Viewer (it only shows the text “Binary data”). However, as before, you can view the xml code with “Raw XML View”.

Conversion

In http://xena.sourceforge.net/media/Text_version-v0.2.pdf , it is described what file formats that Xena can convert to plain text (“Text Normaliser”). As for “normal” normalisation, Xena tries to guess the format of the input file(s).

The following input formats are supported:

- TIFF
- HTML
- Office
- PDF
- Plaintext

When you try to convert a file of a format type that Xena doesn't recognize (for example, Djvu), a warning is given and the file is not converted, only copied to the output directory.

I tried to convert all the 145 files at once (except the very large TIFF file). The only files that Xena managed to convert were the TIFF files (they were converted to PNG). For some reason, an empty version (0 bytes) were also produced for each successfully converted TIFF file.

I also made a conversion of the TIFF files where the option to produce a text version was chosen. Text versions were produced for four out of the six input files (the text versions are *not* readable in a text editor).

4.4.3 About XML output

XML is a very general format, that can be used for almost anything, and thus it may be very useful to have XML versions of your archive files. It is, of course, more useful for some formats than for others; for example, for office documents (“office” in general, not “MS Office”) and other text documents. Therefore, it's a drawback that Xena was not able to normalise/convert Excel documents (and succeeded only partly with the ODT document).

Observe that the XML files cannot be opened in a common browser; apparently, you have to have some Xena metadata files available to do that. These metadata may be available somewhere, but if that is the case, they are not easily found on the Xena web site (see middle paragraph in the next section).

4.4.4 Recommendation

Xena is very easy to use for batch conversion; when you have supplied default input and output directories, it takes very little effort to normalise/convert a lot of files (or at least, to *try* to do so). However, it needs a bit of trial-and-error to understand what actually normalisation, binary normalisation, and conversion is about, and what the results will look like.

If you like XML and have formats that are suitable for conversion (for example, TIFF, jpeg, csv, *maybe* also Open Document Format(s) and MS Office documents), this tool may be useful. For binary conversion, everything(?) seems to work (but you have to figure out how the result can be used later, and how to verify the results). As earlier stated, the results cannot be viewed in a common web browser. Probably, this is because the schema link that can be seen in the “raw xml output”, http://preservation.naa.gov.au/xena/1.0, is out of date.

Xena recognises many more formats than those that have been tested here (see page 15 in http://xena.sourceforge.net/media/How_Xena_ids_file_formats.pdf), but you cannot always trust this. For example, Excel is mentioned as one of the supported formats, and this format was recognised by Xena but impossible to convert. Open Document Format is also supported, but doesn't seem to recognise Swedish characters.

Grade

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

Simplicity of installation: 2 – 3

Simplicity of management: X

Ease of use: 4

Generality of solution: 4

Quality of result: 2