CONVEVAL: ASSESSING THE "CONVERTIBILITY LEVEL" OF AN EXISTING DOCUMENT SET

Andrea Trentini
Dipartimento di Informatica e Comunicazione
Via Comelico 39
20135 MILANO
andrea.trentini@unimi.it

ABSTRACT

This paper presents a methodology and a tool to compute the overall "convertibility level" of a document (file) from proprietary formats to open ones. It is based on the decomposition of the document in its objects (e.g. images, tables, etc.) and on the assignment of "convertibility levels" to different types of objects. This way we can compute a weighted average to (somewhat and a priori) measure the effort and the expected result in the conversion process.

KEYWORDS

FLOSS, integration, conversion, migration

1. INTRODUCTION

This paper presents part of a methodology and a set of tools that we are developing to define and assess the level of "openness" (http://en.wikipedia.org/wiki/Openness) of a cluster of firms. The overall architecture and some implementative details have been described in [TMM07] and [Tre08].

In complex existing business systems (or network of) very large amounts of data are already present or flowing from node to node. Modern firms tend to organize themselves in networks of task oriented delocalized units. The goal of managing complexity drives this trend, but a truly modular organizational setting - where units interact only through predefined transactions - enables parallel work and may improve the flexibility with respect to the uncertainty of the future [BC03]. Intuitively, the nature (format and/or packaging) of data has an impact on the value of the information exchanged. In fact, given the same content, standard and open formats bring more value than proprietary and closed ones. One motive for this is that information value is directly proportional to its accessibility/availability, and open formats are more accessible than proprietary ones.

We developed a methodology, codenamed NorVAL (NetwORk eVALuation) [Tre08], that analyzes many pieces of information (e.g. files and packets) scattered through the network of a group of interconnected firms. The analysis computes an "openness value" based on a weighted average of all the information gathered. The idea is simple to describe:

1. define values (weights) for every class/subclass in the network model (example below)
2. gather information about all the data stored and flowing in the network of firms
3. compute value

For example, speaking about documents (files), we may arbitrarily define a range of values for different types of document formats, e.g.: ODT (0.8), DOC (0.5), OTHER (0.3). Then we gather information about the number of documents in the network, e.g.: ODT (1230), DOC (3500), OTHER (45). So that the value of the TextDocument class in the example network would be: 0.8*1230+0.5*3500+0.3*45 = 2747.5. Thus, for example, by only converting every DOC instance to an ODT instance we may reach a much higher value of 3797.5. Of course the mathematical formula can be changed (e.g. normalized) but the general idea remains

the same.

This level of granularity can of course be refined. This is the object of the present paper.

In fact, not every document is the same: even the simplest MSWord document contains several objects (images, paragraphs, fonts, tags, etc.). When converting to OO.org ODT format these "features" must be mapped into "almost corresponding" features, possibly including information loss. The conversion is usually not perfect and, depending on the set of features used in the original document, the result may vary.

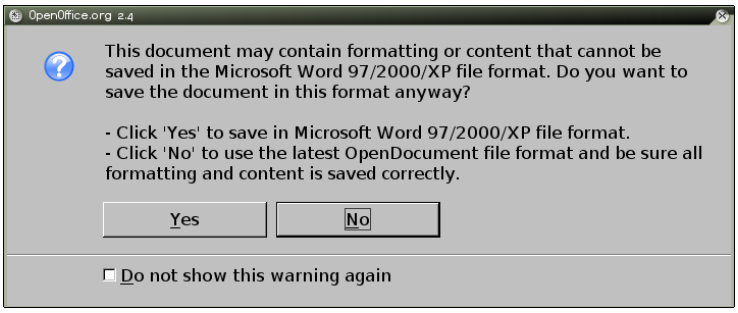

Figure 1: Here's the warning given by OO.org when converting to MSWord, it says that loss of information may occur, of course vice versa can happen the same loss.

So that, if it's true that migrating from a closed format to an open one (e.g. from "MSWord .doc" to "OO.org .odt") raises the overall value of the data, the conversion process may erode some of this value... We'd like to compute this erosion by defining a "convertibility value" computable on a per-file basis.

2. IDEA AND IMPLEMENTATION

The methodology developed in the larger project NorVAL[Tre08] gathers information about files, services and packets. At file/document level the procedure creates a table of the documents available in the network taken into consideration. This table is then weighted, assigning values to the various formats, and a global network value is the computed, as in this example output:

		TOTAL VALUE =>	**18048.2**
Numerosity	*Type*	*Weight*	*Value*
(gathered)	*(gathered)*	*(assigned)*	*(weight*numerosity)*
13585	txt	0.6	8151
5761	png	0.5	2880.5
3101	jpg	0.5	1550.5
2197	gif	0.5	1098.5
1881	xml	0.7	1316.7
1465	html	0.6	879
1353	pdf	0.4	541.2
1203	htm	0.6	721.8
...	...	...	...
282	doc	0.4	112.8
254	rtf	0.3	76.2
248	asp	0.1	24.8
239	sh	0.9	215.1

234	tex	0.9	210.6
...	...	...	...
57	odt	0.8	45.6
...	...	...	...
53	asf	0.3	15.9

Table 1: An example listing of the cumulative value of an existing set of documents. The first column lists the number of files present in that class (second column), the third column lists the given weight (parametrical) and the last column the total.

In *Table 1* you see the overall value, a weighted average, of the net computed for the actual population of documents, where every document of the same format is considered undistinguishably.

The ConvEval tool aims at evaluating the internal structure of every document to analyze the actual components of the document itself, to give **a more precise value to each document instead of assigning "tout court" a single value to the used format.**

So that, for example, we may define of course a weight for a format (like the 0.5 assigned to MsWord DOC in the example above), but this weight can then be adjusted on a single file basis depending on the actual content of each file parsed.

The rationale of a "convertibility value" is that a single file, e.g. an MsWord doc, can of course be converted to an OO.org to raise the "openness value" of the population of data, but the conversion is not perfect and this level of imperfection depends on the types of objects (e.g. tables, images, etc.) contained in the original file.

So we define a "convertibility value" for a single file as:

$$convertibility(file) = (\sum convertibility(o) \; for \; every \; o \in Objects(file)) / |Objects(file)|$$

Where the *convertibility(o)* of a single object must be assigned studying the actual ability in conversion of the currently available software (i.e., we have to compare the original object, say a table, with the converted object in the converted document and "humanly" rate the conversion effectiveness).

We implemented a beta version of ConvEval by fostering the scriptability of OO.org[OO.08]. The tool opens every file in a specified list and outputs the structural content of each document: objects with their numerosity. This information will be used, when we'll assign *convertibility(o)* to each *o*, to compute the *convertibility(file)* value. Here follows an example output (actual figures are fictitious):

```
==========================================
/home/user/.....................FileName1.doc
Number of graphical objects:1
Number of text table:0
Number of embedded objects:0
Number of shapes:4
Number of frames:0
Number of bookmarks:0
Number of reference marks:0
Number of footnotes:0
Number of end notes:0
Number of text sections:0
Number of document index:0
Compatibility factor: 5
==========================================
/home/user/.....................FileName2.doc
Number of graphical objects:11
Number of text table:1
Number of embedded objects:1
Number of shapes:-3 (A bug's always present! )
Number of frames:7
Number of bookmarks:32
Number of reference marks:0
```

```
NUMBER OF FOOTNOTES:5
NUMBER OF END NOTES:0
NUMBER OF TEXT SECTIONS:0
NUMBER OF DOCUMENT INDEX:0
COMPATIBILITY FACTOR: 5
```

We are still analyzing a bunch of converted documents to assign values to single objects in the documents. Basing our study on the frequency of conversion errors measured (alas by hand, visually comparing the original document with the converted one) on a sample set of documents (about a hundred files) we chose to assign the following weights (range [0=no problems]..[5=very problematic]):

- Embedded Object 5
- Text Table 5
- Graphic Object 4
- Frame 4
- Shape 3
- Text Section 3
- Foot Note 2
- End Note 1
- Bookmark 1
- Reference Mark 0.5
- Document Index 0.5

3. CONCLUSION

We developed a tool to analyze the structure of every document in a network, it parses every document and list the structural content in order to assign a "convertibility value" that can be used to adjust the "format value" of the NorVAL methodology.

We are working on the following aspects:

- test the software on a number of documents and assign values to different objects
- extend the analysis to other formats, at the moment we can do it only on formats recognized by OO.org

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial contributions from the FIRB project "International fragmentation of Italian firms. New organizational models and the role of information technologies", a research project funded by the Italian Ministry of Education, University and Research)

REFERENCES

[BC03] C.Y. Baldwin and K.B. Clark. The value, costs and organizational con sequences of modularity, 2003.

[OO.08] OO.org. Uno architecture. http://wiki.services.openoffice.org/wiki/Uno, 2008.

[TMM07] Andrea Trentini, Alessandro Marchetto, and Mattia Monga. Weighing the impact of ICT on "modular" SMEs. In Equity2007, volume ISBN: 978-1-4244-2537-2. IEEE, 2007. Amsterdam.

[Tre08] Andrea Trentini. Norval: a methodology and a system to evaluate the "openness" of a cluster of firms. In Conferenza Italiana sul Software Libero 2008, 2008.