Collaboration and the Knowledge Economy: Issues, Applications, Case Studies
Paul Cunningham and Miriam Cunningham (Eds)
IOS Press, 2008 Amsterdam
ISBN 978–1–58603–924-0

Empirical Evaluation of Network Externalities in Data Standards Diffusion in a Subset of European Organizations

Bruno ROSSI, Barbara RUSSO, Giancarlo SUCCI
Free University of Bolzano-Bozen, Via Sernesi, 1, 39100, Bolzano, Italy
Tel: +39 0471 016956, Email: brossi@unibz.it, brusso@unibz.it, gsucci@unibz.it

Abstract: Network externalities are particular effects in information goods markets that increase the value of goods depending on the number of adopters. This peculiar effect can have relevant consequences in data standards diffusion, a critical aspect of ICT markets due to the impact on interoperability among institutions and software infrastructures. We propose in this paper an approach to help understanding the existence of these effects and evaluating their importance. The approach can help organizational stakeholders to better understand dynamics in different scenarios, like the migration to a new software application or strategies targeted at augmenting the diffusion of a standard inside the organization. We apply the method to a case study of six European Public Organizations focusing on different categories of data standards. We derived useful lessons from the deployment of the approach, above all that the approach - although not fully automated - has been unintrusive in the activities of the organizations involved. In this sense, transparency, and unintrusiveness of the data collection activities are of foremost importance, at least as important as the results submitted to the committing organization. Main findings of the case study are that the presence of network externalities is stronger in some standards categories rather than others, and that a full generalization to other institutions cannot be achieved due to the peculiarities of each organization, from organizational size to budget available.

1. Introduction

Standardization has been defined in [12] as “the process of establishing standards that are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics”.

In fact, standards in the Information & Communication Technology (ICT) sector are considered of paramount importance for the diffusion of knowledge, to hinder the creation of interoperability barriers in e-Government, and to avoid situations of lock-in [2, 11].

In particular, Shapiro & Varian define a situation of lock-in as occurring when consumers are constrained by given past buying decisions due to the presence of high switching costs [11]. A customer will incur high switching costs should he decide to move from a brand of technology to another. They classify lock-in in seven different types, according to the distinguishing switching cost. In this sense, different categories that arise vary from aspects such as contractual commitments, where breaking a contract can lead to compensatory damages, to concepts like brand-specific training, defined as the durable cost involved in training: once training activities have been performed with a specific environment, it may be expensive to switch to another one. In all these cases there is some kind of durable decision that constraints future scenarios.

When considering data standards, switching costs are determined by the large amount of information stored in one format that could hinder possibilities of technological choices.

To better understand this problem, we must introduce another concept: network externalities. This term, was first noted for general industrial markets in [2] as the fact that the individual value given by consumers to a certain good can increase with the number of units sold.

The same argument was later extended to ICT markets, as pointed out in [6]: “the most notable of these characteristics is that software markets often are subject to network effects, whereby the value of a piece of software (e.g., an operating system) rises with the number of other end users who run that same software”.

Typically, a useful classification is between direct and indirect network externalities. In direct network externalities, the value of a good changes for users as a direct consequence of the change in number of successive adopters of the same good. This is the exact phenomenon we described so far. On the other side, indirect network externalities are due to the increase in the number of complementary goods used that has a positive effect on the value of the main good. In this sense, diffusion of data standards connected to software can have a fallback on the selection of the software (like the availability of software for a particular hardware platform tends to increase the value of supporting hardware for potential buyers).

We can see in this way that when considering data standards, the effect can be particularly critical, as the modalities of their diffusion can have large impacts on interoperability among different systems and institutions. A more detailed discussion can be found in particular in [5]. The motivation of such outcome is thus evident: the selection of a data standard that is potentially vendor-dependent will be valued by users not only for the intrinsic value of the standard - and the openness of the standard can play a major role in this consideration - but also for the number of other users or organizations that made the same decision.

In this paper, we empirically evaluate network externalities in the diffusion of data standards in a subset of six public European organizations, raging from small to large organizational size. We first present related studies, objectives, and empirical methodology adopted. We then focus on the technology case description by applying the methodology. Lessons learnt, practical suggestions for replication, main findings from instantiation of the methodology, discussion and conclusions end the paper.

2. Related Studies

As reported in the introduction, mostly of the literature about network externalities has captured interest - among others - due to the seminal works in [2, 6, 11]. Similar studies to the one we propose are the ones related to the empirical investigation of data standards diffusion and general evaluation of network externalities. In this sense, there are not many studies that can be specifically compared to our approach. Among them, we must cite [3], where the author studies the impact of data standards on the PC market. Results corroborate our findings: among different standards examined, one is found to be relevant in order to generate network externalities in the PC software market. In [4] the compatibility with a spreadsheet format is signalled as important for the generation of network externalities. The interest is in both cases the market, while our focus is the adoption process inside single organization. From a technological point of view, similar software agents to analyze file generation process have been proposed with different objectives in dynamical form as in [7, 13] or in static form as in [1, 10]. Our approach is more static as in the last studies mentioned: the data collection process is performed as a one-shot operation and not by continuously monitoring users. What is lost in terms of potential information collected is earned in terms of less invasiveness of the data collection process.

3. Objectives

The main research question of this work is to assess the effect of network externalities in the context of data standards diffusion inside European Public Organizations. It is indeed an important problem to analyze, as these effects can push or augment the usage of standards that otherwise will be only marginally used. This is not usually an issue, but can become problematic in the case of standards that are more open in terms of specifications than others. In this way, even if the strategy of the company is to promote the usage of open standards, the final results can be very different as expected. As seen in the introduction, there is a more subtle problem: the presence of data standards in an organization will constrain future choices of the applications utilized.

Providing an approach to help understanding the existence of these effects and evaluating their importance empirically are thus the main objectives of this study. In order to perform the investigation, we proceed by defining a process of investigation, a methodology of analysis, and then we instantiate the methodology to a case study.

4. Methodology

We use an empirical approach in answering our research question. In the specific, we use an inductive model, concerned with the generalization of theory as the outcome of observed real world phenomena determined throughout the collection of quantitative data. We propose here the high-level process of investigation (Figure 1).

First, we collect information about data standards available in the target organization. From a set of users that typically use several software applications to perform the assigned tasks, we analyze the files that are generated. From the extensions of the files we can then derive the data standards used.

In order to perform the data collection process, we developed a custom application called FLEA (FiLe Extension Analyser), a software application that scans disk drives and catalogues all the files found according to their extension and content. The software agent records different metrics, like extension in the file name, file size, date of creation, and date of last modification.

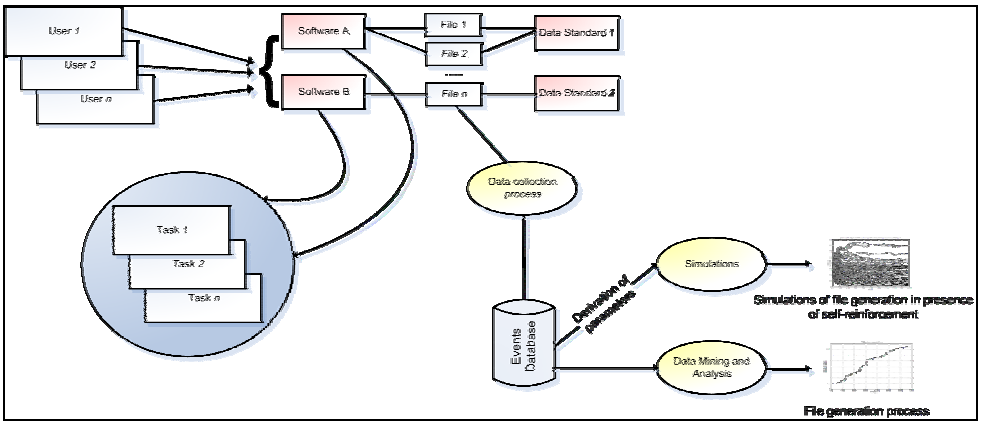


Figure 1: High-level process of investigation

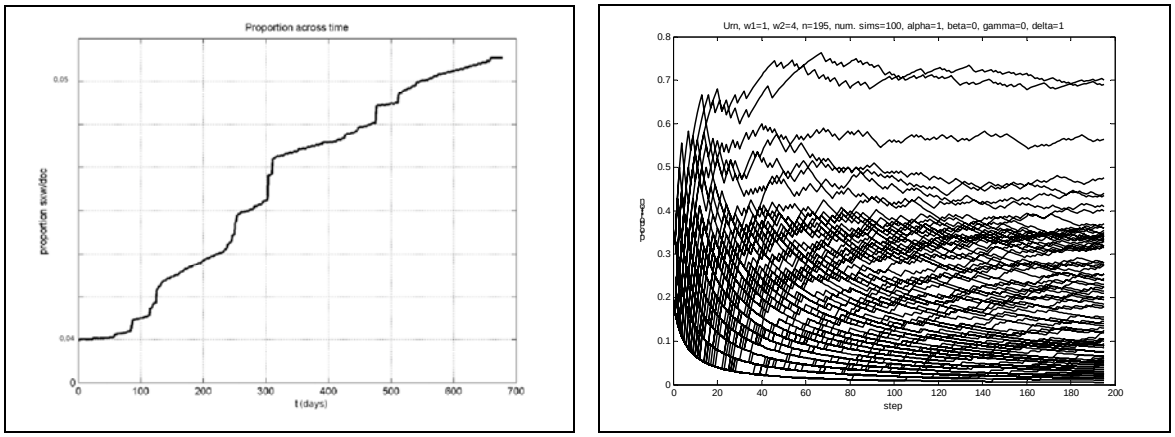
As a second step, we perform an analysis about the diffusion of standards, reconstructing the process of file generation (see, in this sense [8]).

Third, we analyze the evolution of data standards by means of a particular stochastic modelling technique that defines the level of dependence of the file generation process from past history, as presented in more detail with the aid of simulations in [9]. We then compare the results of empirical data collection and the simulations in order to evaluate the degree of fitting between the real world behaviour and the modelled behaviour.

The main idea is that by performing such comparison we can respond to the main objective of our study, that is evaluating the level of importance of network externalities. We provide here more details about second and third aspects of analysis.

4.1 – Analysis Approach

In order to evaluate network externalities in files generation connected to particular standards, we need to consider the self-reinforcement effect of the process of generation. To summarize briefly from the introduction, we consider a process to be self-reinforcing if users will tend to generate more and more files of a certain standard as more files of the same type are available. In this sense, and to better exemplify the problem, we propose our approach to consider file generation in Figure 2.a. At each moment in time (x-axis) we represent the cumulative proportion of files of a particular format that have been created up to that particular date (y-axis). We consider proportion between different standards, as this can better model the spreading effect of different and competing standards. This is the part that considers the real world behaviour.



a) b)

Figure 2: Example of generation of files of standards in time (a), and simulations of an urn process (b)

At this point obviously we do not know whether the behaviour collected for several users and several data standards is subject to network externalities in the process of adoption. For this reason, and in order to evaluate the self-reinforcement of the process, we use a modelling technique based on urn processes. Urn models are particular stochastic instrumentations that allow to model cumulative self-reinforcing mechanisms. We skip here many of the technical details about model definition, parameters estimation, and model fitting. Interested reader can find this information in [8, 9].

To model the process in presence of network externalities, we run different simulations to get a non-deterministic process (Figure 2.b). Also in this case, for each moment in time (x-axis) we represent the cumulative proportion of files (y-axis).

The simulation reports information about the expected behaviour in case of reinforcing mechanisms that limit users' choices in some way. We can see in fact that the behaviour of generation is constrained to some specific outcomes of usage. At this point we have the observed behaviour and the expected behaviour under our study hypothesis.

We need now to compare and evaluate the two processes. To do so, at each point we average the process of file generation and we compare the expected value with the observed value of the file generation process we are examining by using a distance function. To evaluate the correspondence of the observed data with the hypothetical distribution given by the model, we then use a Kolmogorov-Smirnov goodness of fit test.

5. Technology Case Description

We present here the case study that we used as a benchmark for the methodology proposed. Six European Public Administrations were involved in the data collection process.

The Public Administrations that we included in the analysis, were coming from three different countries, namely three from Italy, one from Hungary and two from Ireland. We present a summary of the sample characteristics in Table 1.

Table 1: Details of sample public organizations evaluated

Organization	Nation	Size	Attitude towards ICT Innovation
PA1	Italy	Large	High attitude towards innovation, small budget for ICT innovation in single municipalities
PA2	Italy	Medium/Large	Good attitude towards ICT innovation, limited experience in this sense
PA3	Hungary	Small	Interest in innovation, limited resources available
PA4	Ireland	Medium	High attitude towards innovation, average budget for ICT innovation available
PA5	Ireland	Medium	High attitude towards innovation, average budget for ICT innovation available
PA6	Italy	Large	Limited attitude towards innovation, average budget for ICT available

We can note that organizations examined ranged from small to large organizational size with different budget available for ICT expenses. Most of the organizations were positive towards the introduction of innovation in their ICT department.

5.1 – Data collection

In this study we only considered files and corresponding data standards stored in a central storage location, thus considering client-server architecture. This is in line with modern IT infrastructures that tend to be organized around a central server granting storage services to client machines. Analyzing servers' storage enables the collection of information mostly related to working activities, thus excluding personal files stored on local machines.

We collected information about over six millions files with the FLEA application. We propose two descriptive statistics of our sample that have been proposed in [8]. Table 2 shows the number of files examined, the total size of the files, the total number of different extensions found and the server/partition type in each organization part of our sample.

Table 2: Number of files, total size of files, total extensions, server type per organization

Organization	Total Files	Total Size (*Megabytes*)	Total Extensions	Server type
PA1	5.409.689	653.187	2.074	Linux/Ext3
PA2	92.713	12.007	14	Windows/NTFS
PA3	19.705	1.176	126	Windows/NTFS
PA4	137.487	20.201	324	Windows/NTFS
PA5	31.345	1.167	129	Windows/NTFS
PA6	615.311	145.728	1.058	Windows/NTFS

We mapped all files generated to different standards associated to the file extension. In this sense, we report all extensions that have more than 20 files that have been generated with that particular extension. We further restricted the analysis on a subset of extensions and significant data standards for our analysis (Table 3).

Table 3: Distribution of files in categories according to the sample considered

Extension	Number of files	Extension	Number of files
Text Documents		RAR	347
DOC	3.700.208	TAR	576
PDF	82.315	ZIP	22.590
PS	2.804	Graphic Formats	
RTF	24.610	BMP	128.182
SXW	130.904	GIF	527.289
TXT	281.474	JPEG	408.988
Spreadsheets		PNG	18.586
SXC	8.603	Drawing	
XLS	510.524	DWG	47.728
Presentations		DXF	2.029
PPT	26.351	SXD	138
SXI	215	Database	
Compression		DB	35.194
ACE	13	DBF	121.989
ARJ	1.508	MDA	4.051
GZ	482	MDB	18.994

From the categories examined, we can see how there is a large prevalence of some standards in particular categories (e.g. text documents, spreadsheets). As we have seen in the introduction, in all these cases switching to a different application not supporting fully the previous standard can be problematic. We apply now the methodology as defined in previous sections to two categories of data standards that we consider relevant for empirical investigation.

5.2 – Analysis

We applied our approach to all six public organizations of our sample, focusing on two categories of data standards, namely text documents and graphical formats.

We consider these two categories interesting in the case study for two different reasons. Text processing formats are particularly relevant in the sample considered, as the main activities of employees are related to office automation. Graphical formats on the other side are important part of the usage of applications connected to the web. In this sense, the pattern of usage of this category is peculiar as the creation of graphical files in the sample considered is mostly done indirectly by downloading the content. For each category we present the type of files considered and the fitting for the dataset deriving from each organization. This is the level of fitting of the real world behaviour to the modelled behaviour in presence of network externalities that we discussed in the methodology section. We present the results of the analysis in Table 4.

*Table 4: Fitting of observed file generation process versus model design simulation. Values marked with * are two-tailed significant at level 0.01*

Data Standards	PA1	PA2	PA3	PA4	PA5	PA6
DOC vs (SXW,RTF, PDF, TXT)	0,012*	0,0013*	0,036	0,013*	0,023	0,012*
PNG vs (BMP, JPEG, GIF)	0,52	0,067	0,066	0,24	0,010	0,077

We can derive two main findings from the results of the analysis. First, in mostly of the organizations considered, text-processing categories are more subject to self-reinforcing effects rather than graphical formats usage. One reason can be in fact that while the former category sees a direct involvement in creation by users in office automation, the other category is merely indirectly influenced by users. Second, only in some cases of the text

processing category fitting is significant by performing a Kolmogorov-Smirnov goodness of fit test. This gives more confidence that the distribution of the model identified better fits the observed values with a certain level of significance. So only in some organizations considered, we can actually be confident about the results obtained.

6. Discussion

The methodology proposed can be used to evaluate the importance of network externalities in an organizational setting. We discuss in this section the usefulness and benefits of the method, express some considerations about the replication of the case study and finally summarize lessons learnt from the experience.

6.1 – Benefits and potential impact of the methodology

To better exemplify the benefit of the methodology, we can do one practical example. We suppose a scenario with an organization that wants to perform a migration of a software application to a new software technology. By performing a parallel addition of the new software to ease the transition, management will gain by knowing not only the availability of standards available inside the organization, but also the importance of network externalities in the standards considered. The reason is that in presence of large indirect network externalities, users will be more reluctant to abandon the old technology in favour of the new. Management will benefit from knowing on which applications and standards focus the attention.

6.2 – Considerations about replication

Replication of the experience can be proposed in any organization that is supported by a general repository for users' files. The data collection process is fully automated by means of the software agent that has been used in the case study provided. Anonymity of the data collection process is also granted, as no information that can be used to identify single users is collected. Data of this phase are also automatically inserted into a Database Management System (DBMS) to ease querying of information.

The analysis phase is still manual, in the sense that operations have to be performed in order run scripts in order to gather information needed from the DBMS.

The simulation phase of the models is semi-automated: parameters have to be gathered form the real-world behaviour examined. After this step, the simulation is fully automated by means of an interpreted language script.

The last phase - the comparison phase - is still manual: data from previous phases must be collected manually and inserted in a software application to perform calculation of fitting level.

We note that even by performing a subset of the methodology proposed, namely the data collection process and the analysis phase, still organizations can gather useful information such as data standards diffusion and distribution among users.

6.3 –Lessons learned

Collecting data inside the organizations of the technological case studies provided useful feedback for the data collection process: aspects of transparency, anonymity, unintrusiveness are of key importance. From a technological perspective, we can note that we opted for a one-shot data collection process. This approach revealed to be extremely useful in order to limit the intrusiveness of the data collection process. From a managerial point of view, the approach also provided the opportunity to disseminate the knowledge about the importance of data standards for the ICT infrastructure of the organizations.

7. Conclusions

As seen, indirect network externalities constituted by generation of files connected to data standards can have an impact on the decision to take advantage of a particular software platform. We provide with this research a methodology that can be used by stakeholders to evaluate network externalities connected to standards diffusion. The benefit is to gather information about possible side-effects on the diffusion of software applications that support peculiar standards. We applied the methodology to two different categories of data standards, namely text processor formats and graphical formats. Generation in both cases presents evidence of self-reinforcing effects, whereas the former category seems to be more subject to this kind of effects. This major finding seems to corroborate general theory that states that large amounts of files stored in one format can have a large impact in case of software migration [7].

In any case, the results presented are not heterogeneous across data standards and across organizational boundaries. In this sense, we cannot provide in this way a generalisation of the conclusions. The methodology of analysis proposed needs to be applied on a per-organization basis.

Acknowledgements

This work has been partially supported by COSPA (Consortium for Open Source Software in the Public Administration), EU IST FP6 project nr. 2002-2164.

References

[1] Doucher, J., Bolosky, W. (1999). A Large-Scale Study of File-System Contents, Proceedings of the 1999 ACM Sigmetrics Conference, 59–70.

[2] Economides, N. (1996). The Economics of Networks. International Journal of Industrial Organization, Elsevier, vol. 14(6), pages 673-699, October.

[3] Gandal, N. (1995). Competing Compatibility Standards and Network Externalities in the PC Software Market. The Review of Economics and Statistics, vol. 77, Nov. 1995, pp. 599-608.

[4] Harhoff, D., & Moch, D. (1996). Price Indexes for PC Database Software and the Value of Code Compatibility. ZEW - Zentrum für Europäische Wirtschaftsforschung / Center for European Economic Research.

[5] IDABC (2004). European interoperability framework for pan-European e-government services, Version 1.0, Brussels.

[6]. Katz, M., Shapiro, C. (1998). Antitrust in Software markets. Competition, Convergence, and the Microsoft Monopoly

[7] Roselli, D., Jacob, R.L., Anderson, T.E. (2000). A Comparison of File System Workloads, Proceedings of the USENIX Annual Technical Conference, San Diego, CA.

[8] Rossi, B., Russo, B., Succi, G. (2008). Analysis about the Diffusion of Data Standards inside European Public Organizations, Proceedings of the IEEE Conference on Information & Communication Technologies: from Theory to Applications (ICTTA2008), Workshop e-Government: State & Perspectives, 7-11th April 2008, Damascus, Syria.

[9] Rossi, B., Russo, B., Succi G. (2007). A Method to Measure Software Adoption in Organizations: a Preliminary Study, Proceedings of ISWM Mensura 2007 Conference, 4-8th November 2007, Palma de Majorca, Spain.

[10] Satyanarayanan, M. (1981). A Study of File Sizes and Functional Lifetimes. 8th ACM SOSP, 96-108.

[11] Shapiro, C., & Varian, H. R. (1998). Information Rules: A Strategic Guide to the Network Economy. Harvard Business School Press.

[12] Standardization, Definition. Retrieved February 2008, from https://island.fim.ucla.edu/EABusiness/GiftMembership/Docs/project_documents/Glossary.htm

[13] Vogels, W. (1999). File System Usage in Windows NT 4.0', Proceedings of the Seventeenth Symposium on Operating Systems Principles, 93-109.