Can Web Presence Predict Academic Performance? – The Case of Eötvös University

László Gulyás
Loránd Eötvös University and
Petabyte Research Ltd.
Székesfehérvár, Hungary
gulya@hps.elte.hu

Zsolt Jurányi
Loránd Eötvös University
Székesfehérvár, Hungary
zsolt.juranyi@gmail.com

Sándor Soós
Department of Scientometrics
and Science Policy
Library of Hungariand
Academy of Sciences
Budapest, Hungary
soossand@gmail.com

George Kampis
Loránd Eötvös University and
Petabyte Research Ltd.
Székesfehérvár, Hungary
kampis.george@gmail.com

ABSTRACT

This paper reports the preliminary results of a project that aims at incorporating the analysis of the web presence (content) of research institutions into the scientometric analysis of these research institutions. The problem is to understand and predict the dynamics of academic activity and resource allocation using web presence. The present paper approaches this problem in two parts. First we develop a crawler and an archive of the web contents obtained from academic institutions, and present an early analysis of the records. Second, we use (currently off-line records to analyze the dynamics of resource allocation. Combination of the two parts is an ambition of ongoing work.

The motivation in this study is twofold. First, we strongly believe that independent archiving, indexing and searching of (past) web content is an important task, even with regards to academic web presence. We are particularly interested in studying the dynamics of the "online scientific discourse", based on the assumption that the changing traces of web presence is an important factor that documents the intensity of activity. Second, we maintain that the trend-analysis of scientific activity represents a hitherto unused potential. We illustrate this by a pilot where, using 'offline' longitudinal datasets, we study whether past (i.e. cumulative) success can predict current (and future) activity in academia. Or, in short: do institutions invest and publish in areas where they have been successful? Answer to this question is, we believe, important to understanding and predicting research policies and their changes.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579037.

Categories and Subject Descriptors

H.4 [**World Wide Web**]: Web mining; H.4 [**Information Systems Applications**]: Digital libraries and archives

General Terms

Scientometrics

Keywords

Scientometrics, Web Content Archiving, Prediction of Performance

1. INTRODUCTION

Scientific research is an inherently social process. As such, in modern days, it is more and more exercised and carried out at the scenes of modern social life. That is, scientific activities are increasingly embedded in virtual arenas: Facebook, Twitter, LinkedIn or in novel initiatives trying to create social media dedicated to research and science. Several such initiatives exists, such as Mendeley or Vivo, but perhaps the one with the largest current momentum is ResearchGate. [6] Our belief is that potentially the best way to channel such efforts would be to implement something like the Innovation Accelerator concept put forward in [4, 5, 1]. In short of that, however, we turn our scientific interest to using existing tools and analysing data currently available.

Research already had its online presence even before the advent of new social media. Academic institutions and researchers were among the firsts to create and maintain their own web pages and to regularly publish novel content there. However, classic scientometrics largely omitted the analysis of such activities, focusing on more structured datasets, like journal databases or other bibliometric data.

Recently, we have embarked on a project that aims at incorporating the analysis of the web presence (and content) of research institutions into the scientometric analysis of these research institutions. Our motivation is twofold. First, we strongly believe that independent archiving is important, even with regards to academic web presence. This

may serve several important purposes, including the indexing and searching of (past) web content. (E.g., public scientific statements by researchers and/or academic institutions.)

Online media venues change rapidly. New content and topics emerge and disappear as the joint interest of the community producing and consuming them changes. Archiving online content is generally not solved, so the dynamics of such public discourses is rarely studied.

This forms these basis our second motivation: we maintain that the trend-analysis of scientific activity represents a hitherto unused potential. We are interested in studying the dynamics of the "online scientific discourse". While archiving Internet content is a vast challenge, even for a limited domain like academia, the continuous extraction of topic categories and the archival of them could result in a useful trace of the public discourse and its dynamics over a given period of time.

Besides these primary goals, the project will enable the trend-analysis of scientific activity from a novel perspective and possibly even the prediction of scientific trends. This may have an enormous potential. As an illustration, we present an analysis, using only 'classic' longitudinal datasets of publications and citations.

The usual question, "can investment predict (or imply) success?", may also be reversed. The customary hypothesis answers this original question to the affirmative: more investment is generally believed to yield more success. Using longitudinal datasets of scientific publications and citations, we can ask the reverse question as well. In particular, our current question is this: Does past (i.e. cumulative) success predict current (and future) investment? In other words, is success a good predictor for subsequent resource allocation? Put still differently, do institutions invest and publish in areas where they are successful? Assuming that resource allocation is rational (i.e. that it is worth investing in areas that have proved to bring returns), our question translates as this: do institutions know (or care about) their own research strengths ([2] and [3])?

This paper reports the preliminary results of an ongoing work. The next section describes the material and methods. This is followed by initial results: from basic statistics to some revealing correlations. A discussion of the current status of the project follows, including areas of future works.

2. MATERIALS AND METHODS

Our current ambition is to archive and to mine web content (and presence) of Hungarian academia. This means 500 NIIF institutions (NIIF = National Information Infrastructure Department), 42 research institutes of the Hungarian Academy of Sciences, and 47 higher education entities (universities and polytechnics).

We use modified harvesting techniques originally tested and tried by several earlier national archives, including the internet archiving project of the British Library. We use the Heritrix crawler modified and specially configured for our purposes. We did not consult existing archives but used the modified crawler to download and archive original content. Our hardware configuration was a Dell T710 server (2x4 core Xeon E5520, 48GB RAM, 2TB HDD). In the current study we concentrated on web data we obtained for 48 academic (that is, higher education and Hungarian Academy of Science, HAS) institutions. The lists of these are:

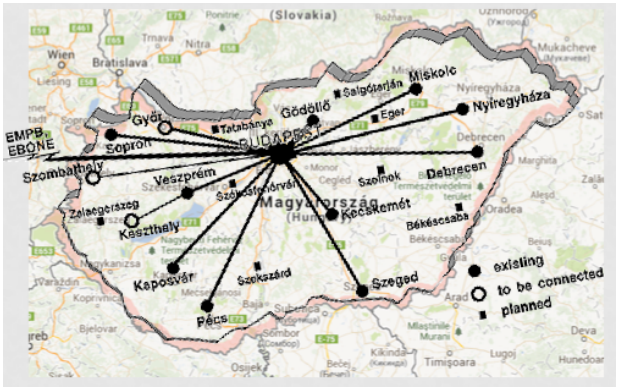

Figure 1: The Hungarian backbone connecting cities of primary academic importance.

- http://mta.hu/mta_kutatointezetei
- http://hu.wikipedia.org/wiki/Magyarországi_egyetemek_listája, and
- http://hu.wikipedia.org/wiki/Magyarországi_főiskolák_listája, respectively.

The list of all these academic institutions is convenienlty summarized at http://www.hungarianscience.org .

Files downloaded are mainly text files and videos stored at the above sites; in particular all files with the extensions exe, gz, iso, jar, mp3, ogg, ppt, rar, wav, xls, xlsx, and zip are excluded (as a response to the existence of many shared disk images and other large files of dubious origin that are often unrelated to the "official" activity of the downloaded sites).

An overview picture of the Hungarian internet backbone serving the sampled institutions is shown on Fig. 1.

In the second part of our analysis, due to the limited time span of our longitudinal data collected from the web at the time of writing, we will deal with off-line recorded data containing metadata of scientific publications contained in the WoS ISI archives. We queried these archives for all Hungarian institutions and obtained a raw dataset containing bibliometric (i.e. publication and citation) data for all publications with at least one Hungarian research address. We used parts of this dataset for the study presented here.

We use these data to analyze the predictability of scientific investments from past success. In this context, success will be measured as citation efficiency, and investment in terms of publication numbers. This can be assumed to be just natural, as publications and the research activities behind them typically tend to imply costs, and the typical return is peer recognition expressed, among other things, in the measurable form of citations by other publications. In short, we deal with (publication, citation) pairs.

Expenditures clearly vary from discipline to discipline and subject to subject. Yet in the same field or subfield the costs can be assumed proportional to the number of publications produced. Hence, normalising publication numbers to their field average, and comparing the results to similarly normalised field-specific citations, we can hope to obtain a clear picture on how the different institutions have allocated their resources at a given time, and how this has led to their success of failure in these fields. We compare numbers about the

past to those of the current investments (the proxy for which is the number of publications) using the same fields. For this study, field are identified by the ISI subject categories (SCs), currently 244 in number in the Reuters Thomson Web of Science (WoS ISI) database, which is the information source of the current study.

We consider the Hungarian dataset (filtered for institutions as listed in http://www.hungarianscience.org/) in the interval 2003-2012.

3. RESULTS

3.1 Hungarian Academic Web Sites

Regarding our efforts to archive the web presence and content of Hungarian academic institutions, our first observation is that the Internet-based "big data" is unexpectedly small for Hungary. The average size of a snapshot of the downloaded sites is 974 MB per site, where the median is 137 MB [!]. (Outliers exists, due to catalogs present in Chemistry sites and astronomy datasets.) The average size of actual text on these sites is 474 MB per site, where the median is 47 MB [!].

By comparison, the text size of the personal page of one of the authors alone is 180MB. When contemplating these figures, then, an inevitable conclusion would be that "big data" is, for the Hungarian academic institutions currently, rather small. This further indicates a lack of tools, competence, interest or available content (or a combination of these) for a more massive web presence of these institutions. (Comparison for institutions from other countries is under way.)

It is well known that averages of skewed distributions should be treated with special caution. Therefore, it is prudent to look at the distributions of downloaded snapshot sizes. The rank distribution of the data sizes of a snapshot of online content for the academic institutions follows a "power-law-like" distribution, as can be expected (Fig. 2). A similar picture is obtained if only text files are considered (Fig. 3).

A quick analysis of the most frequent topics on academic sites in Hungary between April 15, 2013 and October 15, 2013 shows that the focus is on events (*rendezvény*), calls (*felhívás*) and grants (*pályázat*, as presented on Fig. 4. Clearly, the bulk of academic web sites is *not* about the results produced. (A more in-depth analysis of our accumulating data is underway, including spike-analysis, as well as combining the online trends with bibliometric datasets.)

3.2 Prediction from Longitudinal Datasets

In this first report of our analysis, we discuss our next findings on one strongly motivational example, Loránd Eötvös University (ELTE), which is the nation's prime science university. Here we are making use of the off-line bibliometric datasets discussed earlier. Cutting ahead of the detailed analysis, the raw data for 2002 and 2012 can be seen in Fig. 5. This figure — albeit available just for "eyeballing" — already indicates the main finding, namely that there is no obvious relation between past success and later investment as understood here. Notice on the figure that virtually no change is detectable, that is, the successful, highly cited and obscure, not cited fields behave similarly in this longitudinal

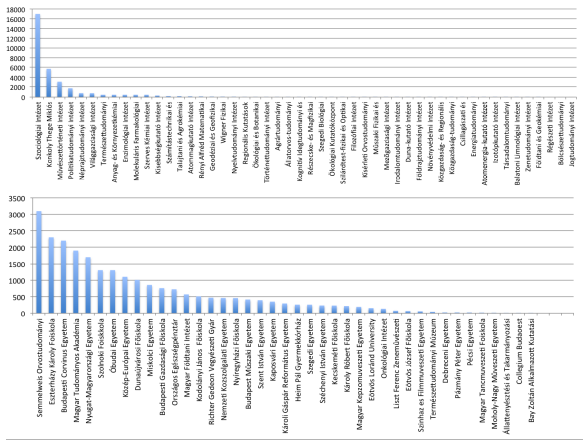

Figure 2: The rank-size distribution of the data sizes of the academic institutions follows a long-tailed distribution. Top panel: Institutes of the Hungarian Academy of Sciences. Bottom panel: Hungarian Universities. In both panels, the horizontal axis lists the institutions from left to right in decreasing rank order. The vertical axis shows the size of the snapshot for the particular institution.

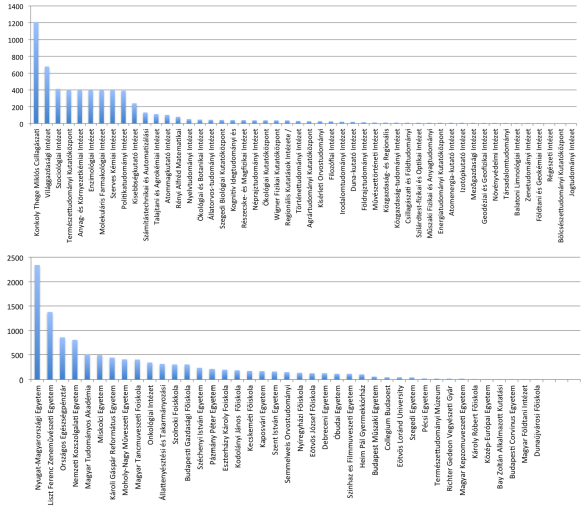

Figure 3: The rank-size distribution of the data sizes of the academic institutions (text files only) also follows a long-tailed distribution. Top panel: Institutes of the Hungarian Academy of Sciences. Bottom panel: Hungarian Universities. In both panels, the horizontal axis lists the institutions from left to right in decreasing rank order. The vertical axis shows the size of the snapshot for the particular institution.

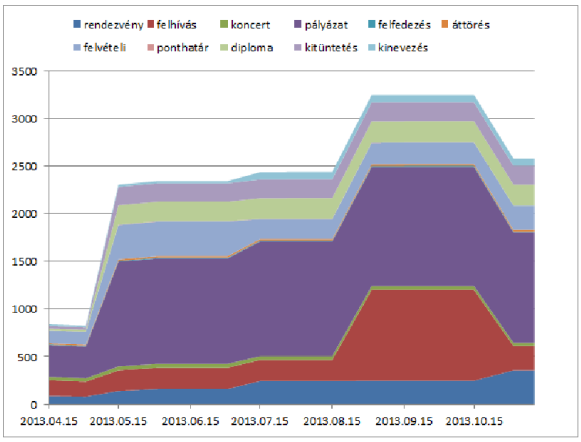

Figure 4: The most frequent topics on Hungarian academic web sites between April 15, 2013 and October 15, 2013. Time is shown on the vertical axis. The horizontal axis show the cumulative number of hits. Different colors denote the different keywords.

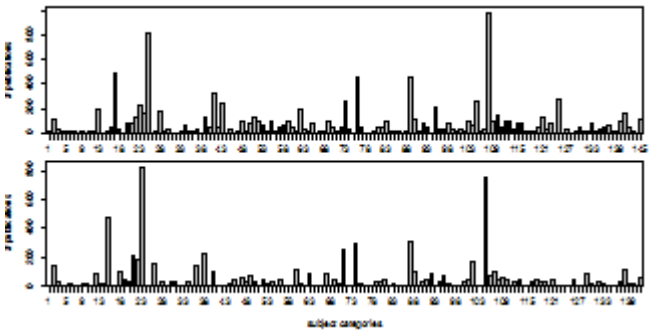

Figure 5: Raw data comparison for Eötvös University, 2002 and 2012 series on the top and bottom panel, respectively. In both panels, the vertical axis lists the ISI subject categories, while the vertical axis depicts the number of occurrences.

comparison (a slightly imperfect SC alignment is due to the small increase of the set of SCs over time).

3.3 Raw investment vs normalized citations

All following figures refer to the same institution, ELTE. The first analysis (Fig. 6) shows raw investment (in 2012) against nationally normalized citations (for papers published in 2002). More precisely, dots represent SCs, the horizontal axis SC-specific relative citations on a national scale, applied to publications in 2002, and the vertical axis shows raw publication numbers in 2012. It is seen that a relation, if exists, is highly unspecific.

Relative citations and nationally normalized citations are obtained as follows (all calculations for a given SC; hence SCs are treated independently). A relative citation number is simply the number of citations divided over by the number

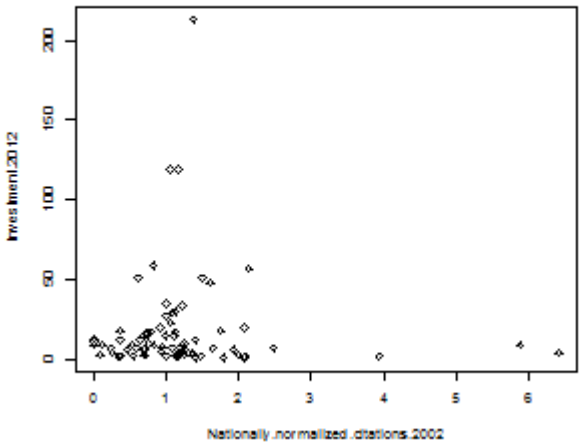

Figure 6: Raw production in 2012 for Eötvös University (verticla axis) versus nationally normalized relative citations (horizontal axis) in 2002.

of publications — for example, if, in 2002, 27 papers have been published by ELTE in Agriculture, and together they received 54 citations, then the relative number of citations is 2. (Note that citations are always cumulative, and so is the ISI database itself — in a current dataset only the total number of citations arrived for the 2002 papers up until 2012 are available. Hence it cannot be known how successful a 2002 paper — or SC — "initially" was. This is not a problem as we are considering a longer interval where initial effects get smoothed out.)

Nationally normalized citations are obtained by first calculating relative citations for the national production (e.g. 207 citations for 73 papers in Agriculture yields 2.84), and then dividing the institution-specific relative citations by the national figure. In our fictitious example, we get $2/2.84 = 0.7$.

What we see in Fig. 6 (if anything), it is that in impact there is a strong grouping around the value of 1 (which cannot be surprising as 1 is, by construction, the average of nationally normalized productivity), and at this value there is a broadest spectrum on investment "responses" on the vertical axis. This indicates that average (i.e. mediocre) success decouples form later investments in the same topics (but see "large fields" later). Even more surprising is that very successful fields (with x-values 4 or 6) are not the ones highly published later.

3.4 Nationally normalized production versus normalized citations

On Fig. 7 we present a similar picture, with the difference that instead of the raw publication numbers, a relative measure showing nationally normalized publications in 2012 are shown on the vertical axis. The finding is similar — that there is no recognizable pattern of relation. Success (or the lack of it) in 2002 and production (or the lack of it) in 2012 are unrelated — a somewhat counterintuitive result.

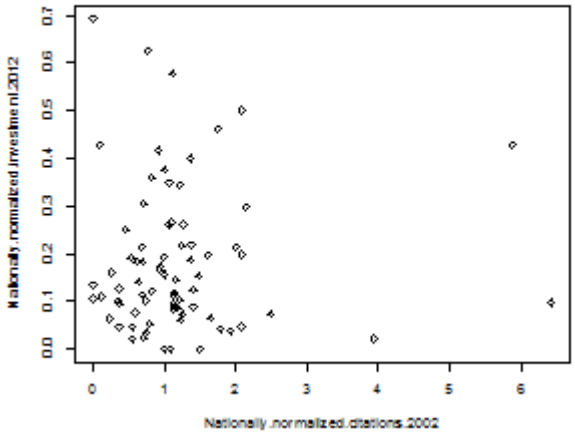

Figure 7: Normalized production in 2012 (vertical axis) versus nationally normalized citations in 2003 (horizontal axis).

3.5 Growth in given fields versus normalized citations

As a sanity check for the ELTE material, we present an analysis figure using a different indicator, the amount of growth in the different fields as dependent variable. Growth of an SC is understood here as the ratio of productivity, the ratio between the numbers of papers produced in 2012 and back in 2002. It is well seen on Fig. 8, however, that ten-years growth in a given field is highly unrelated to the success of the same field ten years earlier.

3.6 Raw citations and later investments: field size effect

Finally let us present an obvious picture of analysis — it is obvious in a double sense. First, it should be clear form the preliminary presentation of Fig. 5 that "not much has changed" in the studied years — our analysis will now express this in more accessible terms. Second, in bibliometrics there is a well documented "field size effect": that large fields remain large, and small fields remain small (so changes, if any, do not cross orders of magnitude, for example). Note that this is conditioning, at the same time, the number of available citations in the given fields: as a general rule, large fields have more citations and small fields have fewer citations. On Fig. 9 below, we see this field effect in the case of raw citations versus later investments.

4. DISCUSSION AND FUTURE WORK

We have reported on an ongoing effort with multiple aims. We are in the process of longitudinally indexing the public web content of Hungarian academic institutions. Our goal is to combine scientometric analysis with trends extracted from online academic presence. These efforts are at an early stage, therefore our current report in this domain was limited to the basic statistics of our downloading and indexing activities, with some illustration of the topic analysis made possible by the former. Nonetheless, important conclusions can be drawn from these basic statistics as well. In partic-

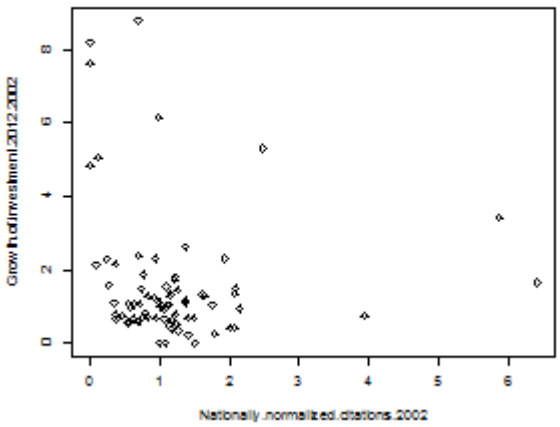

Figure 8: Growth in a decade (2002-2012, vertical axis) versus nationally normalized citations in 2002 (horizontal axis).

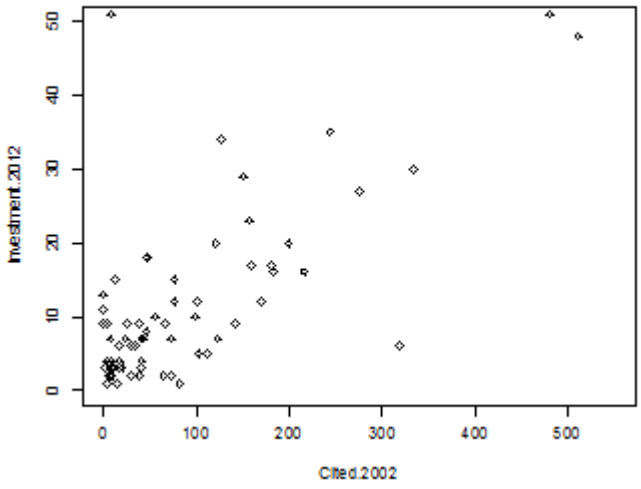

Figure 9: Raw citations (for publications in 2002, horizontal) and later investments (publications in 2012, vertical).

ular, we found that "big data" can be surprisingly small in case of countries similar to Hungary.

In addition, we studied the question of whether scientific success predicts the future (academic) activity of research institutions. This study, performed using 'offline' longitudinal data sets, and considered as a pilot for the combined system (under development) based on web crawling, was intended as an illustration of the potential of trend analysis to predict scientific trends (or the lack of them). We have presented various analyses of one particular example, Eötvös University, in the interval 2003-2012. We found that there is no clearly detectable relation indicating the effect of past success (measured in the number of field normalized citations) on future investment (measured in the number of new publications in a field) ten years later. (An earlier version of this study was presented in [2] and [3]).

Our result indicates that (despite common intuition) the academic institutions — or at least some of them, as revealed in our study — in fact do not know or care about their strength and weaknesses when deciding about future resource allocation in given fields. A very similar result can be obtained if not only Eötvös but also other Hungarian higher education and research institutions are examined. Further discussion of the findings (e.g. whether the unveiled cognitive information strategy represents a strength or a weakness in given cases) is left to subsequent papers. One particular goal we have is to extend this line of research to perform the same set of analyses on data from internationally leading academic institutions (e.g., Harvard University), to see if the results generalize.

For policy and research, these preliminary findings convey a negative message. If institutions allocate their resources (as reflected in the number of papers produced by them) without consideration of the past success of the respective fields, then success does not lead to increased innovation, and may not even be worth wanting, after all. To corroborate these conclusions, and to see whether such a radical conclusion is indeed supported in general is to be found in subsequent works.

Our ultimate goal will be to combine the two sides of the efforts presented in this paper, and to be able to perform scientometric (trend) analysis on combined datasets: i.e., data from web crawling and bibliometric repositories combined. Our belief is that this is within reach and may pave the way for further combined analyses, e.g., ones that overlay funding information (i.e., grants awarded) and publication and impact data, resulting in an approximation of ROI of research investments.

5. ACKNOWLEDGEMENTS

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV- 2012-0013).

6. REFERENCES

[1] G. Kampis, 2013: Innovation Acceleration by Public Data Analysis, presentation at the FuturICT.hu "Networking Conference", Budapest, 14th June 2013.

[2] G. Kampis, L. Gulyás and S. Soós 2012: Megjósolható-e a ráfordítás a sikerből?, Előadás az V. Emergencia Workshopon, Budapest, 2012 december 7.

[3] G. Kampis 2013: Approaches to activity mapping and performace evaluation in scientific production, talk given at the University of Duisburg, Aug 8., 2013.

[4] Helbing, D., Balietti, S. (2011). How to create an innovation accelerator. The European Physical Journal Special Topics, 195(1), 101-136.

[5] van Harmelen, F., Kampis, G., Bőrner, K., van den Besselaar, P., Schultes, E., Goble, C., ... and Helbing, D. (2012). Theoretical and technological building blocks for an innovation accelerator. The European Physical Journal Special Topics, 214(1), 183-214.

[6] Leydesdorff, L., Rotolo, D., and De Nooy, W. (2012). Innovation as a Nonlinear Process, the Scientometric Perspective, and the Specification of an Innovation Opportunities Explorer. Technology Analysis & Strategic Management (Forthcoming).