

INTERNSHIP REPORT

FEBRUARY 2007 – APRIL 2007
(PART TIME, TOTAL OF 30 FULL WORKING DAYS, 5ECTS)

AT THE

VIRTUAL KNOWLEDGE STUDIO

Student	Johannes von Engelhardt (0275964)
Programme	Research Master Social Sciences
UvA supervisor	Prof. Dr. Ed Tan
Practical supervisor	Dr. Anne Beaulieu

Introducing the VKS

The Virtual Knowledge Studio for the Humanities and Social Sciences (VKS) is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), aimed at the reflection and study of the interplay between technology, science and society. In somewhat more concrete terms, the VKS offers a collaborative space for academics from a range of different fields to investigate how the social sciences and humanities make use of new information and communication technologies, how the social sciences can and should do research on and about the Internet and how this challenges traditional ideas about how to “do science”. Presently, ten full members work at the VKS permanently, with a number of visiting and honorary fellows, occasionally participating in joint projects and symposia. The VKS is located in the building of the International Institute of Social History (IISG) in the east of Amsterdam.

While not being a pure research institute, one stream of work done at the VKS is empirical of nature or at least to some extent based on empirical data. Following a Research Master’s Programme in the Social Sciences, this quite naturally was the area in which my work at the VKS was situated.

While my internship involved a range of different tasks and smaller explorations, three general areas/projects can be identified as sticking out as the central themes around which most of my work at the VKS was organized. In the following, I will present these three themes and discuss my role and contribution to each. As will become clear in the course of reading, these three themes are closely interrelated and overlapping and are presented here one by one merely for purpose of clarity and coherence.

The hyperlink: reflections and investigations

The first general theme around which a large chunk of my work at the VKS was organized, was looking into the ways in which hyperlinks are approached theoretically and empirically within the social sciences. My main task here was to map and demarcate the remarkably wide range of theoretical approaches and empirical instruments employed within the social sciences when studying website-linking. Basically, this meant conducting a literature review, identify similarities and parallels, and eventually producing an annotated bibliography (see Appendix A). Looking at how different scholars go about in their investigation of this comparably novel artifact of human action proved to be a particularly interesting and rewarding undertaking for me, for two main reasons. Firstly, my general interest in new media stem from my personal history of working as a freelancer in the field of new media. Secondly, as became quite clear in the beginning of my work at VKS, a discussion of different approaches to hyperlink analysis to some extent inevitably is clustered around the general dichotomy of qualitative/constructivist vs. quantitative/positivist paradigms. In the course of my studies, this dichotomy and its right of existence was a reoccurring theme, and one that I had become rather interested in.

In the case of hyperlink analysis, we can distinguish on the one hand approaches that focus on the counting and statistical analysis of links in order to make claims about large-scale structures on the world wide web. Questions that are being addressed relate to for example who is a relevant online actor for a given topic (Rogers, 2002) or how certain communities are organized online in terms of inter-linking¹. For a review of these and similar largely quantitative approaches, see Park and Thelwall (2003). Within this approach, particular attention is paid to methods of data analysis that are appropriate for the nature of the web, i.e. being in principle non-hierarchical. Here, methods subsumed

¹ Particular attention has been paid here to scientific communities and their member's linking behaviour (e.g., Thelwall, 2003; Barjak & Thelwall, 2007; Vasileiadou & van den Besselaar).

under the term network analysis have been suggested to be the right path to follow. A brief but concise introduction to network analysis and why and how it can be useful in hyperlink analysis is given by Scharnhorst (2003). This type of research is characterized by the use of crawler-software that is sent off to the web (starting off from one or more predefined points of departure) to travel from one hypertext document to the other following inter-document hyperlinks and thus collecting and storing information about link structures. Besides gathering link information, crawlers can also be used to accumulate information about the nodes of the network (i.e. the hyperlink documents) such as domain names (including the Top Level Domain, in many cases identifying the country of origin) or even content-related data.

A number of critical points can be raised against such an approach and while it would exceed the scope of this document to give a comprehensive account of the ongoing methodological discussions, one point will be mentioned that became to dominate my way of working on and thinking about link analysis while being at the VKS.

By quantifying links in the manner described above, we implicitly assume that we can understand them as being equal (or similar enough to be treated as equal).² However, this assumption of equality (or similarity) cannot easily be maintained when looking at the diversity of links on the web. The question that arises is how much we distance ourselves from the object of study and how much we are able to draw relevant conclusions when implicitly assuming that “all links are equal”. This point is also raised by Thelwall (2006) who offers ways of dealing with it without dismissing the whole concept of quantitative link analysis altogether.

On the other side of the spectrum, we find scholars that approach hyperlinks from a school of thought more closely connected to constructivism and qualitative

² Equality here would mainly refer to functional and/or motivational equality, the former referring to the function the link fulfils for the user and the latter to the question of the creator’s motivation to place the link in the first place

methodology. Here, the term “virtual ethnography” coined by Christine Hine (2000) stands central. In these authors, the focus is laid on context and meaning of the hyperlink, thus asking the sort of questions that an “all links are equal” approach by definition is unable to answer, and even more so unable to pose (see e.g., Beaulieu, 2005; Hine, 2007; Beaulieu & Simakova, 2006).

The core themes of discussion between the two perspectives presented that I encountered evolved around issues such as context, generalisability, subjectiveness and polysemy, i.e. those that generally define the well-known front-lines of debate between those who like to count and those who don't. While some efforts of facilitating cross-method thinking and thus countering methodological fundamentalism on both sides have been made (e.g., Howard, 2002), I found that the study of hyperlink analysis constitutes a truly divided field. Having a background mainly in quantitative methodology, this is also one of the reasons, that I very much enjoyed working at the VKS together with Anne Beaulieu who has her “roots” in ethnography, engaging in numerous and fruitful discussions on exactly these issues.

The Athena project

Besides charting out the field of hyperlink analysis, I was also involved in a research project led by Anne Beaulieu. The overarching VKS project under which our research fell, is labelled “Women's Studies and ICT: creating a mediated space of knowledge?” and aims at understanding how the interdisciplinary discipline of women's studies makes use of and thus appropriates new information and communication technologies. Specifically, Anne and me looked into the case of the Advanced Thematic Network in European Women's Studies (ATHENA), an EU-funded network of different women's studies related institutes and organisations, with the main purpose of facilitating international collaboration and communication in both education and research in the

field. Within ATHENA, a number of project websites have been set up for different aims, such as WEAVE for promoting the interaction of young women's studies scholars, or Travelling Concepts for facilitating the discussion about core concepts and themes with the discipline.

Anne had already done some extensive research on the topic and had amongst other things conducted interviews with individuals involved in the different projects to get a better understanding of their use and understanding of the technology employed within ATHENA.. Starting from the results of the literature review on hyperlink analysis mentioned above, I then began to look at the link structures of the ATHENA websites and their different project sites. While it certainly would not be useful to provide a full account of all the analyses I conducted, I will present some of the techniques and results in order to give a general idea of the nature of my inquiries.

In my analyses of the ATHENA websites and project sites, I focussed very much on quantitative hyperlink analysis. The reason for this was firstly that both Anne and me saw it as an interesting challenge to trying to combine these methods with the ethnographic approach Anne had taken in the project. Secondly, as I was and am much more familiar with quantitative methods, this gave me the opportunity to apply some of that knowledge to a new object of study, i.e. the hyperlink.

The first instrument I used for finding and collecting in- and outlinks of the sites under study was google. Furthermore, I used two "shrink-wrapped" tools available for link gathering and link analysis: Richard Roger's Issue Crawler (www.issuecrawler.net) and Touchgraph (www.touchgraph.com). The Issue Crawler is based on the idea of discourse authority. Its point of departure is the assumption that websites that are being linked to by a lot of other websites on a give topic have a greater discourse authority than those that are not being linked to (an idea that is also fundamentally embedded in the google algorithm). It is therefore based on the notion of co-linking (linking to one website by

more than one other website). Touch Graph is a simple tool that is based on google's "related sites" function. Google identifies websites as being related based on amongst other things co-linking, mutual linking and key word co-occurrences. The user provides Touch Graph with a starting website and Touch Graph uses the google engine to identify sites that are being seen as related. Touchpad then proceeds to find related sites to those sites identified in the first place, and so on.

Amongst other things, by using google, I was able to identify some patterns in the inlinking to the ATHENA websites, leading to statements not only on which sites are being linked to the most often, but also on how these inlinks were distributed across different countries (analysing the Top Level Domains of the URL) for the various sites. Thus, some claims could be made about the web "embeddedness" and its international character of the different sites under investigation.

Touchgraph and Issue Crawler were used to look at more complex linking and co-linking structures around the ATHENA websites. To give an impression of the kind of results obtained with these tools, figure 1 shows the graphical output of the Issue Crawler for the ATHENA 2 website.³ From this graph (and the accompanying textual output) it was possible to see which other websites were part of the "discursive network" around the ATHENA site. Again, the aim was to look into the online context of the ATHENA site as defined by linking and co-linking structures.

³ More precisely, a set of inlinks identified by google to the ATHENA 2 site were used as starting points, as the Issue Crawler prompts the user to provide a set of websites to begin the co-linking analysis with.

figure 1
Issue Crawler network around the ATHENA 2 website

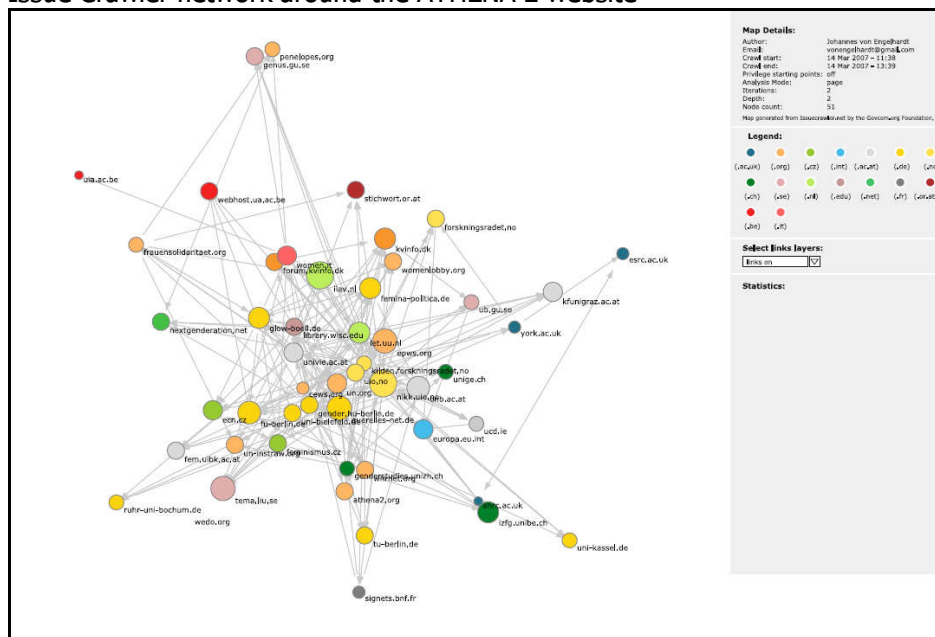


Figure 2 is a visual representation of the network of sites that are “related” to the ATHENA 2 site as produced by Touchgraph. Again, the aim of this largely exploratory method was to acquire some ideas about online relationships, that might not be obvious to the individual website user. One of the findings of the Touchgraph analyses was that certain ATHENA project sites were much more embedded into a network of academic sites while others proved to be more closely related to non-academic sites. Also, in the case of the Issue Crawler, conclusions could be drawn as to the origin of some of the sites forming the network derived from their Top Level Domains – the colours of the nodes in figure 1 represent different Top Level Domains, some of which country-specific. As no concrete research questions were posed at the beginning of these investigations, no concrete, singled-out conclusions can be presented here. The main aim of employing these methods should be more understood as investigative in looking for new sort of questions that could be posed in the future.

possible interpretations, I also addressed some more fundamental methodological issues in the study of hyperlinks, such as the “all links are equal” assumption briefly mentioned above.

One central point of my part of the speech that I will talk about in more detail here was my critique on exactly the type of instruments I had been using in the ATHENA research. This critique was based on the simple fact that we don't *really* understand them. After all, we do not fully know how Google arrives at its set of related sites (and we probably never will as the relevant algorithm is not publicly available). The same is true for the Issue Crawler, which startlingly lacks any in-depth documentation of its inner workings (while producing stunningly beautiful and elegant visual representations that are just so easy to get carried away by). The question thus arises to what degree can we interpret the results produced by a tool that we cannot fully comprehend. As mentioned above, the results of Issue Crawler and Touchgraph (and other easy to use, shrink-wrapped, but poorly documented tools) could yield results that can lead us to ask new questions. However, if we want to be able to give a clear interpretation of the (visual or non-visual) output, it is necessary to understand what is happening: when we are using these highly formalized methods, it is crucial to know the formula. This however, might be difficult in many cases. One way to open up the blackbox could be to use more complex (i.e. more costly) and custom-made tools that fulfil the specific of the researcher. Furthermore, it has been argued that simple search requests can produce valuable (and directly interpretable) results. However, we have to note that for example in the case of google, the achieved results might also be much less straightforward than intuitively expected, due to the fact that we do not know exactly how they are produced either. Anne ended our talk by talking about how we aimed to combine the quantitative techniques I had been using with the ethnographic framework within which the overall project was situated.

Besides the opening speech, Anne and me also participated in the thematic sessions in which a very wide range of researches and research ideas were presented. This proved to be an interesting experience for me, as some of the presented work was quite radically different in terms of theory but mainly method than what I had been used to being confronted to. In this sense, the stay in Barcelona was much more than just a practice in public speaking.

Looking back

In retrospective, I can say that I benefited from my stay at the VKS in a number of respects. Firstly, and most straightforwardly, by reading a lot on hyperlink analysis and related topics, I expanded my knowledge on an issue that I found and find challenging and interesting. Secondly, getting the chance to present at an international workshop was an extremely exciting and instructive experience. Also, as noted above, seeing Internet research being done starting from premises so fundamentally different from my own to some extent broadened my horizon and to some extent also led me to position myself more firmly and consciously with respect to some fundamental questions about the role of theory and the empirical material. Thirdly, by closely working together with Anne, I became much more aware of my own position with respect to the function of theory and method for the social sciences. Through many lively discussions with Anne on both the ATHENA project in particular and on how to conduct social science in general, I believe to have gotten a better understanding of the dichotomy in method mentioned above, of the reason for its existence, but also of how mutual understanding and possible even real (partial) approximations could be achieved.

To conclude, I can state that my internship at the VKS was a rewarding experience and provided me with some new perspectives that I did not come across during my graduate training at the UvA. I also have to stress that my colleagues at the VKS contributed

greatly to making my stay there a very enjoyable one. Particularly, working together with Anne was a true pleasure and her faith in my abilities a real source of motivation.

REFERENCES

- Barjak, F., Li, X. & Thelwall, M. (2007). Which factors explain the web impact of scientists' personal homepages? [Electronic version]. *Journal of the American Society for Information Science and Technology* 58[2], 200-211. Downloaded March 1st 2007 from http://www.scit.wlv.ac.uk/~cm1993/papers/Barjak_Li_Thelwall_preprint.doc
- Beaulieu, A. (2005). Sociable Hyperlinks: an ethnographic approach to connectivity. In C. Hine (ed.), *Virtual Methods: Issues in Social Research on the Internet*, pp. 183-198, Oxford: Berg.
- Beaulieu, A., & Simakova, E. (2006). Textured Connectivity: an ethnographic approach to understanding the timescape of hyperlinks [Electronic version]. *Cybermetrics*, 10[1]. Downloaded March 1st 2007 from <http://www.cindoc.csic.es/cybermetrics/vol10iss1.html>
- Hine, C. (2002). *Virtual ethnography*. London: Sage.
- Hine, C. (2007). Connective ethnography for the exploration of e-science [Electronic version]. *Journal of Computer-Mediated Communication* 12 [2]. Downloaded March 1st 2007 from <http://jcmc.indiana.edu/vol12/issue2/hine.html>
- Howard, P.N. (2002). Network ethnography and the hypermedia organization: new media, new organizations, new methods. *New Media and Society* 4[4], 550-574.
- Park, H. W. & Thelwall, M. (2003). Hyperlink analyses of the world wide web: A review [Electronic version]. *Journal of Computer-Mediated Communication* 8[4]. Downloaded March 1st 2007 from <http://www.ascusc.org/jcmc/vol8/issue4/park.html>
- Rogers, R. (2002). The Issue Crawler: The Makings of Live Social Science on the Web [Electronic version]. *EASST Review*, 21[3]. Downloaded March 1st 2007 from <http://www.easst.net/sept2002.html>
- Scharnhorst, A. (2003). Complex Networks and the Web: Insights from Nonlinear Physics [Electronic version]. *Journal of Computer-Mediated Communication* 8[4]. Downloaded March 1st 2007 from <http://jcmc.indiana.edu/vol8/issue4/scharnhorst.html>
- Thelwall, M. (2003). What is the link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8. Downloaded March 1st 2007 from <http://informationr.net/ir/8-3/paper151.html>
- Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework [Electronic version]. *Journal of the American Society for Information Science and Technology* 57[1], 60-68. Downloaded March 1st 2007 from [\http://www.scit.wlv.ac.uk/~cm1993/papers/Interpreting_SSLAR.pdf

Vasileiadou, E. & van den Besselaar, P. (2006). Linking shallow, linking deep: web use of scientific intermediaries [Electronic version]. *Cybermetrics* 10[1], 61-74.
Downloaded March 1st 2007 from
<http://www.cindoc.csic.es/cybermetrics/articles/v10i1p4.html>