‘Wish you were here before!’
Who gains from collaboration between computer science and social research?

Daphne Duin[1], David King[2], Peter van den Besselaar[1]

[1]VU-University Amsterdam, Dep. of Organization Sciences. & Network Institute, De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands

[2] Department of Computing, The Open University, Milton Keynes, MK7 6AA, United Kingdom

Introduction

This short paper summarizes experiences of interdisciplinary research collaboration by the VU-University Amsterdam and The Open University, in the context of an e-infrastructure called Scratchpads. Scratchpads[1] are an online platform for scientists facilitating e-science in the field of Biodiversity Research.

We understand e-science as having two strands: 1) Collecting, organizing and analyzing often large scale digital data of various kinds, and the tools and infrastructure needed for that; 2) the study of the heterogeneous factors that influence the design, uptake and use of e-science tools and infrastructures. Or in other words, we distinguish between e-science and research of e-science (see also Woolgar, 2004). 2) The second strand uses perspectives from the social sciences (science and technology studies, organizational sciences and anthropology). The two come together when we do e-social research about the use of e-science and e-research infrastructures.

In order to build and implement e-infrastructures like Scratchpads, expertise stemming from computer science and the social sciences are both needed. Obvious contributions from computer sciences will help to build the technical infrastructures for e-science. However, technology alone will not do the trick, as e-science is as also about new ways of doing research, the transformation of work practices, and eventually the changes of the ‘reputational currency’. Consequently the technological design needs to take into consideration various user related issues such as motivation, skills, trust, organizational constraints and incentives, career perspectives, and recognition. Hence, the importance of computer science and social science for e-science. Yet, a third type of expertise is needed which combines the insights from computer science and the social sciences, informing sustainable design and development processes. Collaboration is also important when studying the use of e-infrastructures as done in this paper: using the digital traces of users to describe, analyze and evaluate use patterns.

More generally, working with large or electronic data sets is a challenge for most social scientists. It demands different skills and knowledge compared to the use of traditional social research data, such

[1] Scratchpads are part of ViBRANT, a Research Infrastructure project funded under the 7th Framework programme

as questionnaires, interviews and statistical data. Moreover, social science departments generally lack facilities and support that are needed to work with large or electronic data sets, expertise that is widely available within computer science. Thus e-science has much to gain from the knowledge and skills stemming from the fields of computer science, social science and from the interdisciplinary knowledge of their combined efforts. Moreover, we argue here, that next to e-science, also the individual fields of computer science and social science have much to gain from working together.

In this presentation and in our project we will primarily refer the interdisciplinary contributions from computer science and social science and how they are applied in a project on the development of an alternative evaluation metrics for e-science.

Interdisciplinary insights: Alternative metrics for e-science

"Despite the enormous unfolding investment in e.g. grid technologies, it seems we know almost nothing about how and why (and by whom) these new technologies will be taken up (…)" (Woolgar, 2004 p.2). This lack of methods and theory to evaluate the use of e-infrastructures for science is widely acknowledged and reflects the problems in research evaluation in general. Given the large investments in e-science, evaluating the impact of e-infrastructures is a novel challenge for policy makers and the researchers designing, developing and operating the infrastructure (Horlings et al, 2012). Crucial in such an impact evaluation is the identification of different types of users and use. A variety of alternative approaches to evaluating (e-)science are under development (Duin et al in preparation; Maassen van den Brink et al. 2010)

In our interdisciplinary project, we contribute to this work. More specifically, we use a standard web analytics package (Google Analytics) to generate information on the visitors/users of the e-infrastructure, notably through identification of the names of the visiting Internet Service Providers (ISPs). However, without any further data treatment the names of ISPs have little meaning for evaluation purposes.

The authors of this paper work together on a project towards an alternative metric for (web based) e-infrastructures and as described in the publication "Off-the-rack or made to measure? Developing a metric for e-science in biodiversity research" (currently under review, Duin et al. 2012) they propose: 1) to measure the audiences using (visiting) the e-infrastructure as an evaluation indicator; 2) a method to identify different audience categories.

E-science data

The "e" of e-science stands for 'electronic' and also refers to 'enhancement' (Wouters, 2006). Enhancement because "the core idea of the e-science movement (…) is that knowledge production will be enhanced by the combination of pooled human expertise, data and sources, and computational and visualisation tools" (p 2). Or put differently, the development and application of computational tools and platforms for collaboration and communication in science. These e-science facilities at the same time generate electronic (meta) data, the digital footprints of users and usages

stored in the logs of the e-infrastructure. These data can be used to investigate when and by whom, primary data has been accessed, downloaded, uploaded or edited. In other words, they may inform us about how users 'behave'. The work we discuss in this paper builds on one particular type of digital metadata that we use as data. These data are the digital footprints of users and usages that are stored on (in our case) the server of the e-infrastructure and that can tell us when primary data has been accessed, downloaded, uploaded or edited. In other words, how users 'behaved'. Our work builds on one particular set of use data.

Evaluation of e-science using electronic data

One of the main issues for us, when evaluating the e-infrastructure Scratchpads is the use and use intensity. We studied the visitors and their visit frequencies of the online platform. We were in particular interested in the names of Internet Service Providers (ISPs) coming in to the platform. Visitors of the Scratchpad sites are either registered users or anonymous visitors and so can come from different audience groups, which is reflected in the name of the ISP ('Vrije Universiteit'). Based on the words in the names of the ISPs we clustered them into different user categories (Research& Education; Government; Industry; Media & Arts) with the aim to get insight in the types of audiences using (the information) of Scratchpads. We were in particular interested to see the partition of academic users versus other educational users and sectors such as government and business as this could tell us something about the (societal) impact of the e-infrastructure. The data collection, refinement and classification of the ISPs needed a structured approach. In our project we compared two approaches; the human approach, carried out by the social scientists (DD and PvdB) and a computational approach developed and carried out by the computer scientist (DK) of our group. The computer scientist applied an inductive logic program, aleph and a Bayesian classifier to test a data filtering and clustering method on ISPs visiting the web based e-infrastructure Scratchpads. This collaborative work was formalized in the following question:

> *to what extent can we improve a human developed method with computational techniques, in order to cluster ISPs into meaningful categories representing the various audiences using e-science facilities?*

Additionally, computer science expertise was applied to compare and understand the value of different type of logs files that we could use to collect ISPs data. Initially we had the choice to collect the ISPs from two different web reports: 1) the raw server transaction logs and 2) the visit reports from the web analytics tool Google Analytics, a JavaScript based service. (do we need to discuss this choice in more detail?)

In other words, combining the technical knowledge and skills from computer science with the social science knowledge and skills resulted in a new indicator and efficient tools to measure it.

Results

The first data treatment that was applied to the set of ISPs aimed to remove 'general ISPs' as we cannot link these to an audience category[2].

In our work, we began with the traditional approaches of each discipline to interpreting the data. The social scientists brought their domain knowledge to bear, while the computer scientist applied computational and statistical methods to the data. The effects of these two differing approaches can be seen in the results from our first task, to go beyond Google Analytics report and filter out irrelevant ISPs. The social scientists produced a 181-term filter set after many hours of effort that gave 94% accuracy, whereas the computer scientist produced a 6-term filter set in a couple of minutes that gave 84% accuracy. The tested computer-aided filtering reached a higher *precision* than the manually-developed filter (98% vs 92%) though for the *recall* in this initial test favored the manual approach (73% vs 97%). However, these results need further explanation and in so doing will highlight a key benefit to the computer scientist of the collaboration.

The results are an artifact of the nature of our data. In our gold standard data set of 1000 entries, there are 1,576 unique terms in the names of the ISPs. The most common term is 'of', which occurs 126 times, and the second most common term is 'university', which occurs 118 times. However, there are 1,180 terms that occur only once in the data. Thus, a few terms could accurately identify many of the ISPs, however, to cover the 'long tail' requires a domain expert to review the ISPs because if a term is used only once in the data then there is no pattern of use that a computer can extract. Our collaboration has given the computer scientist a rich data set with which to explore the boundary between domain expertise and the limits of computational and statistical methods.

The boundary between expertise and pattern recognition was explored further in our second task, namely to classify the selected ISPs into their audience groups. We are working at three tiers of grouping: Sector (e.g. research & education, government, industry; media & arts); Level (e.g. secondary education, higher education; and local, regional, national, supra national); and Focus (e.g. water management, biodiversity research, geology, humanities, agriculture, public health). The computer scientist expected the classification to be relatively straightforward after the work devoted to filtering the ISPs and so initially used a simple classifier. In the first run of the simple classifier, we achieved 90% accuracy with Sector and 88% accuracy with Level, both of which are fairly direct classifications. However, the more abstract concepts embodied in Focus, led to an accuracy of only

[2] Some ISPs are commercial companies and providing Internet access is their core business. These ISPs, mainly telecom and cable companies, provide access from peoples' home, or from mobile devices. On the other hand, many companies, universities and government agencies, and non–governmental organizations act as an ISP for their employees or membership. Through the name of the ISP, (i.e. 'Vrije Universiteit') we may be able to identify the nature and activities of the users. When connecting to the Internet through a computer network of the organization, web analytics packages will pick up the name and add it to their 'visitors report'. Hence, web reports contain two types of ISPs: 'general names' of commercial ISPs and 'specific names' of organizations acting as ISP. In the first case the name of the ISP does not tell much about the affiliation of the user, in the second case the name will give a good indication of the users affiliation.

49%. This led to much interesting exploration of how to refine the classifier and the role of more sophisticated techniques such as Bayesian networks.

The filter sets are not directly equivalent in that the more extensive filter set of the social scientists includes redundant terms, ie an ISP could be identified by many of the 181 terms in the filter set. The computer generated filter set does not have such redundancy, once a rule was generated that correctly identified an ISP that ISP was removed from the rule generating process. Arguably, this redundancy means that the 181-term filter set is more robust and will cope better with future data. The rule generation process could be changed but was left unchanged in our work to demonstrate the dramatic difference in numbers of terms identified.

The computer-generated approach can identify patterns not apparent even to the domain experts. For example, ISP names are often that of the institution and a significant pattern related to ISP names emerged in our work, if 'of precedes of' in an ISP name then that ISP should be included in our analysis as it is a relevant ISP, eg 'institute of marine biology of crete'.

One advantage of the computational approach is that everything is recorded, not only the results of the classification but the explanation of how ISPs were classified is recorded in log files. This combination of results and logs allows for a thorough review, and consequent refinement, of the processes. In our collaboration, analysis of the results in Excel using the built-in automatic filters, proved useful not in reviewing the process but in gaining an understanding the data itself as it was classified and re-classified, which led to further refinement of the audience groupings themselves.

The findings of our study emphasize the value of computational techniques for data marking over human data marking in efficiently and precision. Most importantly, the classifier showed that that within minutes of downloading the data from Google Analytics, we could classify with 90% accuracy the "sector" of the ISP.

Based on this experience we argue the use data available on e-infrastructures is a rich source of knowledge for social scientists studying behavior in e-science, but the nature of the data demands for (tailor made) computational tools and data treatment. In many cases e-science data demands for larger storage space, more powerful computers and network facilities, different data security and privacy issues, types of expertise which is for many social scientists are not part of their daily tool kit nor (or limited) part of their training.

Who gains?

Interdisciplinary research helps to solve problems whose solutions are beyond the scope of a single discipline or area of research practice and broader application increases the societal relevance. E-infrastructures are such a domain where computer science and social science collaboration contribute to the development, use of smart technologies that are used and sustained. Hence society in general and e-science in particular may benefit from this collaboration. We argue here that in addition, also the two collaborating fields of computer science and social science have something to gain for their own field and research.

Traditionally social science data are things as: interviews transcripts, survey lists, ethnographic material such a field notes and photographs. The collection of these data sets demand for a physical presence of the researcher and the research object during the data collection phase (except for secondary data analysis such as document or archive research). On the other hand, in infrastructures behavioral information, digital footprints of users are available in the use logs or online forums and do not require physical proximity or direct interaction.

In the context of e-science it is first of all the data collection and preparation of electronic data and large data sets that is new for social scientist. However, also the interpretation of the electronic data poses a challenge for social scientists where computer science insights can contribute, for instance, through helping to identify the meaning of a 'web visit' in a particular web analytics service. This computer science definition of a 'web visit' could then be combined with the 'traditional' social scientists enquiry of what it means to a computer user to visit a website. The size and potential fuzziness of electronic data sets require computational methods to prepare the data before they can be analyzed, automation of data cleaning lessens the chance of human error. In short, electronic data are different in size of the files, in fuzziness of content and its meaning is (partly) defined by technological systems and people's relation to technology, all key issues that can be better addressed if computer science and social science combine forces. Thus, computer science expertise is valuable in the *collection* and *interpretation* of behavioral e-science data.

Computer science is always looking for data to help them test and develop new computational methods and models. This collaboration gave access to an especially rich data set that highlights both the possibilities of computational and statistical techniques and their limitations. The digital footprints in e-science, together with the social science use of the data make it an interesting setting where computer science can make a difference and demonstrate that its value reaches beyond the engineering work of the technical infrastructure itself. In our work, the application of computer science to analysis of the data suggests that the work of the social scientists can be considerably speeded up as through the use of tools to support the initial analysis. This will permit the social scientists to apply their expertise to the more complicates issues only. This approach can be considered with confidence because all work is recorded permitting a detailed review of the analysis. In our work, this review led to improvement not only in the filtering and classifying processes, but to improvement in understanding the audience groups too. Thus, informing and improving the rigor of the final outputs.

Above we listed several reasons for collaboration between e-science facilities, computer science and social sciences, nevertheless every collaboration does have costs: it requires time in planning and communication. Furthermore, collaborators support each other's work often at the costs advancing their own research (cf. Brooks, 1996).

Summarizing, e-infrastructures benefit from interdisciplinary effort of computer scientists and social scientists when it comes to questions of design, uptake and users. Additionally, social scientists benefit as they get new data for their studies. Computer scientists benefit as they have use cases for the tools and techniques they develop, use cases that may test their ideas and help to improve those.

Reference

Brooks FPR (1996) The computer scientists as toolsmith II. Communications Of The ACM 39, 3, 61-68.

Duin D, King DJ, Van den Besselaar P (2012). Off-the-rack or made to measure? Developing a metric for e-science in biodiversity research" (Under review).

Horlings E, Gurney T, Somers A, Van den Besselaar P (2012) the societal footprint of big science. A literature review. Den Haag: Rathenau Instituut. Available: http://www.rathenau.nl/uploads/tx_tferathenau/Working_paper_Economic_and_social_footprintof_big_science.pdf . Accessed 2012 Feb 21.

Maasen van den Brink H, de Haas M, van den Heuvel J, Spaapen J, Elsen J, Westenbrink R, Van den Besselaar P,. *Evaluating the societal benefits of academic research, a guide*. Den Haag: Rathenau Instituut

Woolgar S (2004) Social Shaping Perspectives on e-Science and e-Social Science: the case for research support. A consultative study for the Economic and Social Research Council (ESRC) http://www.ncess.ac.uk/research/social_shaping/oess/publications/2004_woolgar_SocialShapingPerspectives.pdf

Wouters P (2006) What is the matter with e-Science? – thinking aloud about informatisation in knowledge creation. http://www.pantaneto.co.uk/issue23/wouters.htm