What is the matter with e-Science? – thinking aloud about informatisation in knowledge creation

Paul Wouters[1]

The Virtual Knowledge Studio for the Humanities and Social Sciences (VKS)

Royal Netherlands Academy of Arts and Sciences

Amsterdam, NL

Mail Address:

PO Box 95110

1090 HC Amsterdam

The Netherlands

Visiting address:

Joan Muyskenweg 25

1096 CJ Amsterdam, NL

T 3120 4628654

F 3120 6658013

http://www.virtualknowledgestudio.nl

http://www.darenet.nl/en/page/language.view/keur.page

Introduction

E-Science is the buzzword in science policy and science funding. But what does it mean? What sense can we make of it? This paper is actually not much more than an exercise in thinking aloud about how we might get a hold on the phenomenon of e-science.

E-science is generally defined as the combination of three different developments: the sharing of computational resources, distributed access to massive datasets, and the use of digital platforms for collaboration and communication. E stands not in the first place for "electronic" but for "enhancement". The core idea of the e-science movement (most of it is still promise rather than practice) is that knowledge production will be enhanced by the combination of pooled human expertise, data and sources, and computational and visualisation tools. E-science has become a buzzword for funding large-scale facilities, especially in those research fields in which research is driven by huge high-tech research groups.

The stakes are huge because e-science claims to be no less than a revolution in the way knowledge can be created. The domains that are supposed to be affected by this revolution vary according to the national variety of e-science. But taken together the claim seems to be that almost all disciplines will be transformed in some way or other. In the US, the focus seems to be especially on computer science, natural sciences and engineering, and on business applications. They do not talk so much about e-science as about Grid computing applications. "In the next few years, the Grid will provide the fundamental infrastructure not only for e-Science, but also for e-Business, e-Government, e-Science and e-Life." (Berman, Fox, & Hey, 2003, p. 11) The UK has an ambitious national programme comprising these elements, but adding a serious e-social science infrastructure and funding programmes (Hey & Trefethen, 2002). In the Netherlands, things are of course much more frugal. There are some ambitious projects around the physics community, which is at the core of the Dutch Grid community. Here we see a concentration on computer science, life sciences with some links to behavioural sciences. The Dutch case is mainly interesting because the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Information and Communication technology (ICT) foundation of Dutch universities (SURF), formulated the ambition to promote "e-humanities" (Kircz, 2004; KNAW, 2004). Taken together, all disciplines of the sciences have now been included in the ambitions of e-science.

E-Science is supposed to be a new practice of knowledge creation because the new information and communication technologies will create new possibilities. Email and the Web have given only a glimpse of what is possible. "To more fully support the e-Scientist, the next generation of technology will need to be much richer, more flexible and much more easy to use" (De Roure, Jennings, & Shadbolt, 2003, p. 437). So e-Science is for a big part about creating new communication, computing and information

gathering tools. Many texts and projects are basically computer science engineering projects. It is however not limited to the infrastructure. The scientists will have to be transformed. They will have to create new ways of working. They will have to develop a far deeper understanding of what technologically is possible. New breeds of engineer-scientists will be created of which the bioinformatics people are only the first generation. Funding agencies and science managers will also be forced to think in new terms about their business. If they do not succeed in this, the penalty will be missed opportunities in creating new knowledge and technologies. E-Science is as much a message from certain parts of the scientific communities, with large-scale physics at its center, to science managers, as a message from science managers to scientists. And last but not least, e-science is about changing relationships between academia and business. A new era of intimate collaboration seems to be imminent. No longer can academics be allowed to be distasteful about cooperation with industry. Worse, without industry, it is no longer possible to create the high-end, state-of-the-art tools that researchers now need. IBM is the biggest player in the Grid, the e-life sciences are full of smart small companies following the lead biotechnology had given in this area.

Future scenarios

The future practice of e-scientists is usually seen as an extension of present practices, but with added automation, an increased role of autonomous agents, and with a strong emphasis on "coordinated resource sharing", which is actually the main issue in Grid computing. An interesting example of a "motivating scenario" is the following (De Roure et al., 2003, pp. 440-441):

The sample arrives for analysis with an ID number. The technician logs it into the database and the information about the sample appears (it had been entered remotely when the sample was taken). The appropriate settings are confirmed and the sample is placed with the others going to the analyser (a piece of laboratory equipment). The analyser runs automatically and the output of the analysis is stored together with a record of the parameters and laboratory conditions at the time of the analysis.

The analysis is automatically brought to the attention of the company scientist who routinely inspects analysis results such as these. The scientist reviews the results from their remote office and decides the sample needs further investigation. They request a booking to use the High Resolution Analyser and the system presents configurations for previous runs on similar samples; given this previous experience the scientist selects appropriate parameters. Prior to the booking, the sample is taken to the analyser and the equipment recognizes the sample identification. The sample is placed in the equipment which configures appropriately, the door is locked and the experiment is monitored by a

technician by live video then left to run overnight; the video is also recorded, along with live data from the equipment. The scientist is sent a URL to the results.

Later the scientist looks at the results, and, intrigued, decides to replay the analyser run, navigating the video and associated information. They then press the query button and the system summarises previous related analyses reported internally and externally; and recommends other scientists who have published work in this area. The scientist finds that their results appear to be unique.

The scientist requests an agenda item at the next research videoconference and publishes the experimental information for access by their colleagues (only) in preparation for the meeting. The meeting decides to make the analysis available for the wider community to look at, so the scientist then logs the analysis and associated metadata into an international database and provides some covering information. Its provenance is recorded. The availability of the new information prompts other automatic processing and a number of databases are updated; some processing of this new information occurs.

Various scientists who had expressed interest in samples or analyses fitting this description are notified automatically. One of them decides to run a simulation to see if they can model the sample, using remote resources and visualizing the result locally. The simulation involves the use of a problem-solving environment (PSE) within which to assemble a range of components to explore the issue and questions that arise for the scientist. The parameters and results of the simulations are made available via the public database. Another scientist adds annotation to the published information.

The e-science community tends to use this type of future scenario to both motivate other scientists and science managers to support their endeavour, and to identify key technical challenges to create the technological infrastructure that can support these future practices. By staying rather close to what is presently technically possible they both avoid the stigma of science fiction and are able to operationalise the technical challenges into real-world computer science problems. In a more far-away future the latter would become more difficult.

This way of creating future scenarios of scientific practice is, of course, not new. In 1997, De Jong and Rip published a scenario (De Jong & Rip, 1997), that looked a bit further into the future. The key intellectual concept in their paper is the "computer-supported discovery environment" which is at the core of "the computer revolution in science". Although they do not use the concept of e-science as such, they do refer to the same type of expectation and technological scenario:

"The promise of artificial intelligence and other branches of computer science is to radically transform conventional discovery environments by equipping scientists with a range of powerful computer tools including large-scale, shared knowledge bases and discovery programs. We will describe the future computer-supported discovery environments that may result, and illustrate by means of a realistic scenario how scientists come to new discoveries in these environments. In order to make the step from the current generation of discovery tools to computer-supported discovery environments like the one presented in the scenario, developers should realize that such environments are large-scale sociotechnical systems. They should not just focus on isolated computer programs, but also pay attention to the question how these programs will be used and maintained by scientists in research practices."

In their analysis of the discovery environment, De Jong and Rip mention all elements that are also crucial in present-day e-science descriptions and projects: it is a distributed globally integrated system of heterogeneous components that can be seamlessly accessed from any workstation. Most tools and databases are shared, human intelligence is heavily supported by blackboxed machine intelligence and smart agents, and robots do part of the research work. Recently, renewed attention was drawn to the "robot scientist" (King et al., 2004) by creating a new tool that is more powerful than the first robot scientist created in 1990 (Zytkow, Zhu, & Hussam, 1990). This robot is able to originate hypotheses to explain observations, devise experiments to test these hypotheses, physically run the experiments using another robot, interpret the results to falsify hypotheses inconsistent with the data, and to repeat this cycle.

What are the key issues in e-science?

These types of scenarios and visions of the future can be used to tease out the key issues in the e-science developments and movement (in many ways it is a socio-industrial movement). The scenario by De Jong and Rip emphasize the unravelling of new properties of substances and the working of nature. They conceptualise knowledge creation as *computational discovery*. This is still, I believe, the dominant perspective in e-science projects which leads to a strong emphasis on computational science combined with the creation of huge databases with specialised search tools. This is exemplified in the Grid projects in combinatorial chemistry, in the Virtual Observatory built by the astronomy community, in the data-intensive Data Grids for high-energy physics. In these e-science projects the technical research question is usually a well-defined, bounded, computational problem.

However, this is of course only one variety of knowledge creation. If e-science is restricted to computational science in sensu strictu, it could not play the role of a grand vision of the future of

knowledge across the board. Nevertheless, we can expect that e-science will exert a pressure to generate more computational-oriented research also in the social sciences and humanities. E-science also encompasses a different type of knowledge creation, *comparative research*, which is related to the discovery mode. For example, comparison plays a large role in astronomy to discover new phenomena in the universe. Yet, it is different because the emphasis is not on complicated analyses (which are at the heart of for example e-chemistry) but on relatively simple comparisons across huge data sets of possibly very heterogeneous types. The data sets are the key issue. "Data are the killer application of e-science." (Berman et al., 2003) For the technological infrastructure, the size of the data sets is the big challenge. Computer scientists working on these problems focus especially on bandwidth, security and one-stop shop authentication, new types of database management (both able to deal with the size and the heterogeneity of the data) and storage issues. Examples of applications are the UK breast cancer e-science project eDiamond and the applications figuring in visions about the future of the humanities and qualitative social sciences. In the latter cases, the idea of large-scale comparison is not always focused on scientific data in the traditional sense, but may also include other types of resources that can be queried in computer-assisted environments. In some applications, the Grid is seen as a smart interface to distributed resources, this is the case in the so-called Access Grids. The latter may exemplify the third figure of e-science, *seamless access to digital resources*. In the first instance this is more external to the process of knowledge creation, although the way researchers process the resources might be affected fundamentally.

To sum up, three issues are recurring in e-science projects: developing automated computational sciences (also in areas that were not computationally oriented in the past) – *knowledge creation as computational discovery*; processing huge and heterogeneous data sets – *knowledge creation as comparison of data*; seamless access to distributed and heterogeneous digital resources – *knowledge creation as reading in a digital library*.

This threesome should be supplemented with a core characteristic that comes up in most, if not all, e-science projects: *large-scale technologically mediated collaboration*. The e-scientist of the future is no individual human, nor an individual cyborg, but a node in a network of humans and machines. She may be a Nobel Prize winning node, but a node nevertheless. Collaboration here means different things that challenge deep-rooted ways of working in many fields. It means sharing of research data instead of jealously guarding them. It means submitting oneself to standardized ways of working instead of the idiosyncratic behaviour many researchers have now. It means putting time and energy in constant communication instead of closing oneself off to be able, for example, to write a book. This challenge to the sociology of research is seen as the biggest hurdle for the e-science paradigm, not only with respect to the social sciences and humanities, but also in many natural sciences. It is the basis for the market for sociological "Science, Technology and Society" (STS) research directed to e-science developers as

well as to the scientific communities at large. How will quality control be affected? In what ways will the reputational control mechanisms have to be changed? What does publication and authorship mean in the context of e-science? How does one share data that are randomly distributed with no apparent owner anymore? Funky stuff for STS people, urgent questions for the e-science community.

Not a unified phenomenon

This paper is not the place to discuss this program which is being developed in our STS community. I wish to focus here on the question whether it is useful to try to analyse e-science as somehow one phenomenon or a related set of movements.

Of course, e-science is not a unified phenomenon at all. Very different practices and technologies are being put together. The drive to unify them is a significant move at the interface of science policy and the need for new technological infrastructures in certain parts of the scientific community. It is no coincidence that physicists-entrepreneurs are at the core of most of e-science. E-science is a discursive construction at the interface of technical-scientific practices, computer technology design and science policy.

To analyse a phenomenon means also to try to take critical distance. This is the reason to be a bit suspicious if we as STS researchers take over the unifying categories from the communities we study. This is also the danger of this paper, by the way. A unifying dimension of e-science is the infrastructure of information and communication technology that is needed to make the e-science dreams come true. Therefore, any STS analysis that takes the infrastructural technology or the development of scientific instruments as the core perspective on e-science will almost inevitably, (although: what is inevitable?), end up reifying e-science as a significant unified phenomenon indeed. This seems to me for example the case with Nentwich's -in many ways exemplary - analysis of cyberscience (Nentwich, 2003). The book is a very ambitious discussion of everything related to the use of the Internet in academic and scholarly research. Thereby it reifies at the level of phenomena a process called "cyberscience" as distinct from traditional science. It is moreover based on an impact model of technology assessment, which brings the author to conclude that, with variations across disciplines, cybertools will more and more influence the process of knowledge creation. This will lead to a virtualisation and dematerialisation of academia, although this process will not be complete due to resistance of some academic institutions (for example the need for face-to-face contact). This conclusion is however as much built-in in the conceptual framework as in the empirical materials collected. "Cyberscience" is a hypothesis which also holds for e-science itself. In line with the sceptical tradition in STS, it seems therefore sensible to be careful in unifying e-science too quickly (Woolgar, 2002). This conclusion is nothing new, but seems

worth repeating nevertheless.

Therefore, the obvious solution would be to take the opposite approach and focus on scientific practices at the level of the individual researcher, research group or institution. This may also be the level where most questions zoom in on the STS market for e-science mentioned above. The challenge then is to include the networked nature of present-day scientific practices. For this, the inclusion of either social network analysis (Wellman & Berkowitz, 1989) or virtual ethnography (Hine, 2000) seems to be appropriate. This promises to be a fruitful approach and has already been taken up with respect to e-science. It adds critical distance to the ICT-centred technology assessment work. The consequence, however, is that the unity of e-science disappears. E-science is then only one dimension of the scientific practices that we study. What remains is a technological infrastructure that can be coupled to the local research context and practice in a variety of ways. Studies of this type might tend to emphasise the dominance of the local practices as the shaping forces of the development of new forms of knowledge creation. By their perspective they may tend to neglect the influences of other localities than the research locale under study. This form of myopia may seem a lesser risk than the one of reification, but I am not so sure about this. Splitting can be as misleading as lumping. Perhaps it is possible to mobilise other STS resources to study whether or not the use of e-science tools has systematic implications for local research practices? It is in this respect that I think Rheinberger's notion of epistemic objects (Rheinberger, 1997) may be useful, especially if it can be coupled to the emerging methodology of virtual ethnography and Internet research.

On digital representation

To understand e-science, it seems useful to first define the process of informatisation in science. To quote from our earlier research programme *Networked Research and Digital Information* (Wouters, Beaulieu, Park, & Scharnhorst, 2002):

> *"Our working hypothesis is that scientific and scholarly research practices are undergoing an informational and communicational turn[2]. In a growing number of areas, digital information has started to take on a new function within research. The emergence of digital information, embedded in " ICT", has enabled a radical lowering of the costs related to information dissemination. At the same time, new research technologies have affected the process of data generation itself. They have enabled new types of experiments (sequencing technologies), measurements (astronomy), imaging (medical sciences) and data visualisation (modelling software). These have in their turn vastly increased the level of data production in research. Where this happens, scientific research is becoming more dependent on information and communication technologies. Not only information seems to play new roles in scientific knowledge practices, this also holds for communication. It should be noted that the catchword 'internet' actually denotes a number of different modes and technologies of communication: email, Web sites, chat, database interfaces, on line conferencing etc. An important and often*

overlooked consequence of this multimedia feature is that these different modes are all represented in the same digital domain ("cyberspace"). One may therefore better speak of "converging media" as a basic trend. This convergence enables new forms of hybrids of communication modes that were quite separate before the Internet era. It moreover also enables hybrids between communication and other research practices, as can be seen on Web sites where conferences, databases and email and chat facilities are seamlessly merged in one scientific resource."

Epistemic objects and experimental systems

We can take this one step further with the help of the approach by Rheinberger which is in many ways a building upon, and extension of, early work in STS on the laboratory (Latour & Woolgar, 1986). Rheinberger distinguishes two elements of experimental systems. The first is the research object, which is called the "epistemic object" that embodies that which is not yet known. The second element is the set of experimental conditions in which the research objects are embedded, which he calls the "technical objects". Within a particular experimental system both types of elements "are engaged in a non-trivial interplay, intercalation, and interconversion, both in time and in space. The technical conditions determine the realm of possible representations of an epistemic thing; and sufficiently stabilized epistemic things turn into the technical repertoire of the experimental arrangement" (Rheinberger, 1997, p. 29). So what is the difference between a technical object and an epistemic one? It is functional rather than structural; there is no "essence" here. "We cannot once and for all draw such a distinction between different components of a system. Whether an object functions as an epistemic or a technical entity depends on the place or "node" it occupies in the experimental context." (Rheinberger, 1997, p. 30). Nevertheless, in every experimental system scientists are quite aware of what is what. The main functional criterion is that epistemic objects are generators of questions. A technical product is stabilised and is first and foremost "an answering machine". "In contrast an epistemic object is first and foremost a question-generating machine" (Rheinberger, 1997, p. 32).

This difference between a technical object and an epistemic one becomes crucial because Rheinberger puts representation at the heart of the scientific enterprise. Not in the sense of a "true representation" but in the sense of systems of signifiers. This makes his approach especially relevant for our aim to understand the interaction between the Internet and knowledge creation: we are thinking (itself a system of re-representation) about the interaction between two systems of representation. "Basically, my argument is that anything represented, any referent, as soon as we try to get hold of it, and, concomitantly, as soon as we try to shift it from subsidiary to focal awareness, is itself turned into a representation. Engaging in the production of epistemic things means engaging in the potentially endless production of traces, where the place of the referent is always already occupied by another trace." (Rheinberger, 1997, p. 104). Since, moreover, there is no representation without a chain of

representations, "the activity of scientific representation is to be conceived as a process without 'referent' and without assignable 'origins'. (...) As paradoxical as it may sound, this is precisely the condition of the often-touted objectivity of science and of its peculiar historicity as well. If we accept this statement, any possibility of a deterministic referential account of science, be it based on nature or on society, is excluded." (Rheinberger, 1997, p. 105)

So what role do these epistemic objects play in the spaces of representation that are created in scientific activities? They are close to Latour's "immutable objects". "What is significant about representation qua inscription is that things can be re-presented outside their original and local context and inserted into other contexts. It is this kind of representation that matters." (Rheinberger, 1997, p. 106) Thinking about e-science, this is especially interesting because the very core of what many e-science projects aim for is the decontextualization of objects and subsequent decontextualization on the fly and in any context. How is this being made possible? By metadata, a rather dull word for information that should describe the "meaning" of the object/data in such a way that other machines and humans can make use of those objects/data in contexts that might have been unthinkable at the moment of the production of the object/data. Metadata are representations of the original context of epistemic objects in such terms that new contexts can be created for these objects to generate new questions. The main trick that should do this work is not simply querying the epistemic object in its new context, but basically the reconfiguring of new epistemic objects by the recombination of already existing ones in new contexts (or in Rheinberger's terms new technical objects). This means of course that also technical objects can be turned into epistemic objects and the other way around. Seen in this perspective, the whole e-science business is nothing new at all, except for the scale and ease with which objects can be interchanged since they are already digital representations. By representing the whole universe of relevant stuff (both technical and epistemic objects) in digital objects, the interconversion is indeed seamless (apart from the hard work behind the scenes and the hard work of producing the material conditions for the whole business of digital representation). But then again: isn't scale and ease all that matters?

To repeat, representation is not related to the idea of mirroring nature, but to the concept of spaces of signification, to the graphematic articulation that is at the heart of the experimental work according to Rheinberger. The notion of inscription is not simply the product of the experimental apparatus; it is really the scientific object that is nothing more than a bundle of inscriptions (Rheinberger, 1997, p. 111). Inter alea, this means of course that reality is nothing more than an attribute of representations, but this is in the context of e-science probably no longer a threatening statement, even for hard-nosed experimentalists.

Implications

So what does this imply for the study of e-science? Think again about the promise as set out in the future scenario. What we saw, and what we keep seeing in almost all e-science projects, is the drawing together of stuff that was previously separated in one domain to which the researcher – as a node in a socio-technical network, has immediate access and to which she adds as the main result of her work. So the product of research is not the stand-alone publication in some prestigious journal or book, but the addition of both technical and epistemic objects to this space of representation (for example in the form of the uploading into technical databases). Of course, this is no "essence" of e-science, but itself a representation using the lens of Rheinberger's approach. Nevertheless, it does give both a way of speaking about e-science which is sufficiently different from 'actor's speak" to be interesting (for them and for us), and a way of formulating a research agenda that has the potential of critically interrogating the very notion of e-science (Woolgar, 1988), while at the same time studying it "in vivo/silico". To sum up: the circulating references might be the most interesting research site (epistemic object) for STS with respect to e-science, rather than specific instances of it.

References

Berman, F., Fox, G., & Hey, T. (2003). The Grid: past, present, future. In F. Berman, G. Fox & T. Hey (Eds.), *Grid Computing. Making the Global Infrastructure a Reality* (pp. 9-50). Chichester, West-Sussex, UK: John Wiley & Sons.

De Jong, H., & Rip, A. (1997). The computer revolution in science: steps towards the realization of computer-supported discovery environments. *Artificial Intelligence, 91*(2), 225-256.

De Roure, D., Jennings, N. R., & Shadbolt, N. R. (2003). The Semantic Grid: a future e-Science infrastructure. In F. Berman, G. Fox & T. Hey (Eds.), *Grid Computing. Making the Global Infrastructure a Reality* (pp. 437-470). Chichester, West-Sussex, UK: John Wiley & Sons.

Hey, T., & Trefethen, A. E. (2002). The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems, 18*(8), 1017-1031.

Hine, C. (2000). *Virtual ethnography*. London: Sage.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., et al. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist, *Nature* (Vol. 427, pp. 247-252).

Kircz, J. (2004). *E-based Humanities and E-humanities on a SURF platform*. Amsterdam: KRA Publishing research.

KNAW. (2004). Developing e-science in the humanities. Amsterdam: KNAW.

Latour, B., & Woolgar, S. (1986). *Laboratory Life: The Construction of Scientific Facts* (2nd ed.): Princeton University Press.

Nentwich, M. (2003). *Cyberscience. Research in the Age of the Internet*. Vienna: Austrian Academy of Sciences Press.

Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*: Stanford University Press.

Wellman, B., & Berkowitz, S. D. (1989). *Social structures: A network approach*. New York: Cambridge University Press.

Woolgar, S. (1988). *Science: the very idea*. London: Tavistock.

Woolgar, S. (2002). *Virtual Society? Technology, Cyberbole, Reality*. Oxford: Oxford University Press.

Wouters, P., Beaulieu, A., Park, H., & Scharnhorst, A. (2002, 2002). *Knowledge production in the new digital networks. Research Programme.*, from http://www.niwi.knaw.nl/en/nerdi2/research_programme/new/toonplaatje

Zytkow, J. M., Zhu, J., & Hussam, A. (1990). *Automated discovery in a chemistry laboratory.* Paper presented at the 8th National Conference on Artificial Intelligence (AAAI-1990).

An early version of this article was presented at the joint 4S & EASST Conference, "Public proofs, science, technology and democracy", 25-28 August 2004, Ecole des Mines, Paris.

Top of Page

Issue 23: July 2006

The Pantaneto Forum Home Page

---

[1] I'd like to acknowledge the contributions to my thinking about e-science by my colleagues at the Virtual Knowledge Studio for the Humanities and Social Sciences and my former colleagues at Networked Research and Digital Information.

[2] *The notion of "turn" has been chosen to indicate that an important transformation seems to be taking place, while we do not assume that the sciences will change in every aspect. Hence we do not speak of an informational revolution.*