

e-Nabling Data: Potential impacts on data, methods and expertise

Dr Samuelle Carlson, Dr Ben Anderson

Chimera, University of Essex

benander@essex.ac.uk

Abstract. Partly as a result of financial inducement (research funds) but also for sound methodological and substantive reasons social scientists in the UK are beginning to engage with the wider programme of ‘e-Science’. The greater computing capacity and larger sets of data to compare enabled by computing, service and data grids also give the prospect that new scientific questions can be asked – those questions which can only be addressed through massive analysis or the federation of disparate datasets. Partly in response to these initiatives we have been studying the nature of scientific collaboration and knowledge building in three case studies. This paper presents preliminary findings from these studies and focuses in particular on the potential impact of ‘e-enabling’ on social science data, methods and expertise.

Introduction

Partly as a result of financial inducement (research funds) but also for sound methodological and substantive reasons social scientists in the UK are beginning to engage with the wider programme of ‘e-Science’. Investments in e-Science technologies are motivated by a multiplicity of factors. First is the urgency to manage the increasingly large quantities of complex data produced by digital technologies and digitally enabled science. ‘Deluge’, ‘waves’, ‘knowledge overload’ are some of the terms used to describe the situation. Another related factor is the concern of funding bodies to ‘repurpose’ their investments to avoid what is in turn termed ‘data tombs in mono-disciplinary silos’ and to see a maximum return on their investments. The greater computing capacity and larger sets of data to compare enabled by computing, service and data grids also give the prospect that new scientific questions can be asked – those questions which can only be addressed through massive analysis or the federation of disparate datasets.

Partly in response to these initiatives we have been studying the nature of scientific collaboration and knowledge building in four case studies. As Thomas’ *Entangled Objects* (Thomas 1991). *Entangled Objects* highlights the diversity and complexity of exchange practices and multiple relationships to objects entangled in a wider regional and global system. Similarly, one of our goals is to document qualitative diversity and complexity in e-(social) science practices which, analogously, takes place in a national and international context of ‘capacity building’.

Amongst other objectives, e-Science aims to make more and ‘clearer’ data available to a wider user base for primary and secondary analysis. As data producers, archivists and curators have always known this requires not only that the data to be presented in specific forms to be circulated, understood and re-used but that the constituent disciplines see such an endeavor as

both practically possible and potentially useful. Key recent studies in the field of eScience such as the NCeSS funded OeSS and Disclosure Risk Assessment projects (see also Jirotko, 2005; Purdam et al., 2005) underline how most of the obstacles to such data provision are less technological than social, ethical, legal, and institutional.

This paper supports this view although we claim that some barriers might primarily be practical or formal. These issues intervene early in the research cycle, at the point of data collection and distribution, before questions of collaboration and IPR can be addressed. As we will show through a comparison of four academic projects from this perspective certain disciplines might appear to be better equipped than others for the uptake of eScience. However, further exploration shows that formal and practical obstacles are present to different degrees in most science practice.

Study method

The four projects we chose vary considerably in their approaches to data collection, disciplinary backgrounds and use of technologies. The four are described in more detail in the appendix but in brief they were:

SkyProject – a national collaboration of scientists focused on the use of data from a few key observation sites,

SurveyProject – a large scale quantitative social science survey data collection project which has data re-use enshrined in its objectives,

CurationProject – a collaboration between an anthropology department and an anthropological museum,

AnthroProject – a long term research project driven by two anthropologists and implemented through a series of research student projects.

The projects we chose to study allow us to compare eScience projects (in the narrow sense of the term), such as SkyProject, with more ‘traditional’ uses of eTechnologies and to question the potential of the latter for scaling up. With representatives of ‘hard’ and ‘soft’ sciences in SkyProject and Curation/AnthroProject – with SurveyProject holding an interesting intermediary position – we could also raise the question as to whether the similarities and divergences found during the comparative exercise corresponded to a quantitative/qualitative divide. They also provide us with a reasonable spread of disciplines across the typology introduced by Becher (1987). Thus SkyProject represents Becher’s ‘hard-pure’ group, SurveyProject represents ‘soft-applied’ whilst CurationProject and AnthroProject represent ‘soft-pure’. Not represented here is the ‘hard-applied’ group although there are more than sufficient projects from this area in the e-Science programme and some of them have already been studied by others such as Hine (2005) and Fry (2006).

In order to compare the four sites we followed the research process cycle to observe the ways in which new technologies impact this cycle and how this represents different challenges for different disciplines. This partly determined whom we were interested in interviewing, for it required us to meet people involved in data collection, processing, analysis and reuse, when these were not the same person. Otherwise we followed a process frequent in ethnography, which is to follow initial informants’ advice on whom to interview next, according to their understanding of the project and to their network. In all we interviewed sixteen persons, often on several occasions, across the four projects together with three additional respondents at the

UK Data Archive at the University of Essex (UKDA) and other institutions. The data were transcribed and coded using ATLAS.ti as a resource for rapid re-analysis and data enquiry.

In addition we carried out participatory observation at a number of UK e-Social Science and e-Science events such as two of the e-Science All Hands Meetings/conferences and the first International e-Social Science Conference. We also followed up and interviewed contacts recommended by the four projects such as further members of the UKDA.

The paper draws on a diversity of sources including informants' descriptions of projects and how they came to be formulated during one-to-one interviews; the reading of internal and external documents produced, including project websites; and observation of interactions within teams over several days and in the case of SkyProject project the analysis of records of online collaborations through jabber, emails, a wiki, and skype.

Results and Discussion

Inevitably such an open-ended research study generated a wide range of insights and results. Rather than provide a shallow but broad overview of results, in the remainder of this paper we focus on just the potential implications of e-Science for the nature of data, methods and expertise.

Born digital data

We start by showing how the difference between data that is 'born digital' and that which is 'multimaterial legacy' constitutes a serious challenge. If it is not an obstacle for physical disciplines that now produce born digital data, it is a major issue for those using or producing heterogeneous or historical materials.

Whilst projects such as SkyProject and indeed SurveyProject automatically create digital material this is by no means the case in CurationProject and AnthroProject. Here for example these might include recordings or transcripts of songs, dances, tales, and daily practices from which similarly heterogeneous records will be produced: letters, diaries, notes, books, photographs, films and material objects.

It is not at all clear how to 'archive' collections of artefacts in anything other than a physical manner although CurationProject has been experimenting with different methods of opening up the performances of collections via representations and databases which allow the source communities to annotate the cataloguing databases.

However the effort required to even catalogue physical materials in such databases is high and those digitised first are often those which do not present this problematic heterogeneity. Within qualitative research, text is most frequently retained. Verbal accounts, primarily make their way to digital archives, and especially those recorded through controlled and coherent formats. The underlying concern that surfaced during interviews with social scientists is thus that, ultimately, eScience might force qualitative materials into quantitative forms and logics (or favour the quantitative side of qualitative studies) and thereby jeopardize both the specificity of qualitative approaches and the means by which quantitative approaches document their context.

Digitisation is also extremely dull. One respondent in CurationProject said 'The data entry is done in a cold room with no light. It's very boring: you need to check the dates and plug numbers such as 1996, 1997, 1998 and so on. Cleaning the database is time-consuming: check

consistency in formats for names; spelling mistakes etc' and . Indeed our study showed that the AnthroProject team spends most of its time transforming the format of existing data and maintaining them rather than creating/collecting new data. As it turns out this is common across all four of our case studies. A significant part of the work to be done on any of the data in all of the projects we studied was its serial, and occasionally parallel transformation, documentation and maintenance.

Newly digitized data will often not substitute for former types of data but may only be *different* and *complementary* kinds of data. To take CurationProject as an extreme case, as one of its curators noticed that

‘For people, the digital images [of objects] become the objects themselves; in fact they are object in themselves. What goes on the web (the object itself never goes out) are just the *performances* of that object.’

Although he promotes the database as being a new collection in itself, a collection of digital images, a specific aura remains with the artefacts. In the same way, the database doesn't fully replace former types of records which remain 'the' authority. In sum, digital technologies might sometimes only add up to existing ones and at considerable extra cost. As one CurationProject member said

‘The old catalogue cards has a tangible nature. It is also the only *true* one. It has an authoritative nature because so many data has been transferred that errors came in and that we need to go back to it. It's different for the different versions of the database. There is no descriptive quality in data: data is data. It's different for the old catalogue cards that gives insightful info other than data' .

Another CurationProject member remarked:

‘[The head curator] likes the manual catalogue cards and also asks people to look at it rather than the e-one: the e-one is good for identifying, but serious research would imply looking at the cards catalogue for hand-writing etc. and people prefer the feel of it [...] If new info arrives, it's not put on the catalogue cards but we keep on adding cards for new collections’.

This has an impact on the status of digital data which, because they complement and redirect users to artifacts or legacy records located elsewhere, could be seen to hold the status of *information* rather than *data* Thomson (2004).

Explicit and implicit forms of knowledge

For collected materials to become data that can be used and reused they then need to be rendered *disseminatable* - to be rendered both *transportable* in concise abstract forms and *intelligible*. As Strathern (2005) has pointed out communication entails a necessary reductionism and this is embedded within the practice of data representation. The manner in which the different projects address this problem differ.

Quantitative disciplines such as SkyProject and SurveyProject appear particularly well equipped with tools for the visualization and representation of the objects and processes they study.

For SurveyProject this involves the extensive re-naming/codeing of variables and decisions on coding frames for verbatim responses. From the point of collection, answers formulated by respondents in words will be visualized and circulated in the form of numbers. ‘Having all the data, we transform them again because people, most analysts, want a much flatter structure’

commented the computer manager. 'Derived variables, weighting, imputations are put on' (ibid.).

What was first in CAPI and passed in SPSS and SIR is made then accessible again through SPSS, SAS and STATA software. Variables names are kept consistent across years as well as the terms used for indexing. It is all this process of conversion of words into numbers; of successive flattening and restructuring; of renaming and renumbering that transform materials collected into visible, manageable, communicatable and intelligible data for its community of users.

Quantitative disciplines not only benefit from numbers as means of visualization but also from mathematical and symbolic systems as means of communication. Their methods are formally and informally taught and therefore shared within communities of researchers. SurveyProject seminars are precisely dedicated to testing the reliability of results by testing methods against others; there are 'final results', 'errors', 'explanations', 'causalities and effects'. As a senior research officer of SurveyProject explained quite neatly

'You can lie with statistics, but not to another statistician, the truth comes from the statistical methods, not the numbers. You will get found out because the numbers and the methods of interpretation are there side by side. (...) But a lot of people don't get the maths: our clients certainly don't. They just want the legitimacy of numbers without the uncertainties of the full model (...) "the numbers" are not self-legitimizing, the full model is what gives you the grounds for legitimacy'.

Although SurveyProject gathers researchers with different backgrounds (economy, sociology, political sciences) who use and name statistical methods differently, statistics nevertheless provide the common language that allows them to expose their methods to another.

If particularly heavy to circulate, the materials collected in SkyProject are nonetheless similarly conveyed in highly visual and stable forms: images, catalogues of numerical values and models. Because the object of study is entirely numerically mapped, an image can be translated in numbers which can be translated in diagrams. To this visualization 'tool-box', one can add images created out of theory. On the issue of visualization, we were told by SkyProject members: 'The raw data is tables; not images. You can do things with numbers; you'll still have to have numbers out of an image.'

Quantitative disciplines thus appear particularly well equipped with tools for the visualization and simulation of the objects and processes they study. Jens Erik Fenstad (1995) observed the same processes at play in Natural Sciences, which depend on 'non-trivial physics and extensive mathematical modelling, e.g. raw data have to be transferred to a standard brain format and enhanced through a delicate centre-of-mass algorithms before they become the reported data in the analysis of the cognitive task' (1995: 63-64).

At first sight this represents a challenge for many qualitative approaches which appear more overtly tacit and craft-like in their operations. AnthroProject does not visualize data in a condensed and abstract form that would easily lend itself to circulation and thus seems to belong to these disciplines for which eScience would imply a radical epistemological change.

In the main anthropologists in AnthroProject go alone, rather than in teams, to the field, and the information they record and the methods they use are rarely externalized until they take the final form of published monographs or articles. In short, most data and the tacit knowledge used to construct it, remains in the head of the anthropologist until s/he formulates descriptions and interpretations in writing. Except for some types of diagrams inherited from the 19th century when anthropology precisely aimed to be a quantitative science the

anthropologists studied appear to make very little use of symbolic visual representations to model the social processes they study. Thus funding bodies' promotion, not only of eScience, but also of interdisciplinarity, increasingly pressures such disciplines 'to give *form* to middle range materials between the largely private archive of fieldwork data and the forms in which ethnography is reported in monographs, articles and even informal papers.' (Brenneis, Marcus, 2005).

Because it seems to represent a limit of eScience, we can push further our exploration as to the reasons why anthropology, but also other disciplines in the Humanities, doesn't fully embrace the prospect of eScience. First, not only anthropology deals with materials that are not easily made visual in condensed standardized forms, but there might also be wrong assumptions on what constitute data for them. Several of our respondents emphasized that what anthropologists are most interested in in each other's work is not the 'raw data' collected from the fields but interpretations, concepts and analyses. To schematize, a pool of primary and intermediary anthropological materials exclusively might not necessarily be of appeal to them. Also, contrary to many studies in quantitative sociology aiming at representativeness and for which there is a lot to gain to compare or compile a greater number of sources, anthropology is about comparing and contextualizing a few number of sources and cases and, likewise, the capabilities offered by eScience might not be necessarily relevant.

Qualitative approaches are often said to rely on tacit knowledge which rests in implicit personal or institutional practices often associated with craft-like skills, awareness of reputations and hands-on techniques. However on closer inspection we found as have others since Latour & Woolgar (1979) that tacit knowledge and craft-like practices pervade all of our case studies. In many senses the quantitative and qualitative divide transforms into a dimension of specificity. Some disciplines are simply more specific, and make more use of external representations than others. Those that do, we argue, will not only be better placed to take advantage of e-Science approaches and the funding that supports them but their data will also tend to be selected for archival and re-use in preference to others.

Disconnecting data from people

Data and claims

Having disconnected data from people in order to make it in some sense 'public' and re-usable we have to remember that social scientists study people. Unlike objects, people have claims. In the case of AnthroProject and CurationProject many data (artefacts) are entangled with the practice and knowledge of multiple others who might consider this data sensitive and whose conceptions of ownership might vary. Of course similar issues now arise in the context of biomedical data where there is not only privacy to consider but also litigious and commercial interest (Coopmans, 2006). This is best illustrated in our case studies by CurationProject who face daily dilemmas over what to make public and in what form.

One of the constraints on the implementation of digital technologies at CurationProject is that any people from one of the many contexts of an object could potentially have a claim on the use made of it. For a given photograph, for instance these may include those depicted, the photographer, the collector or the museum. CurationProject deals with complex data because these data are entangled with the practice and knowledge of others, who might consider it sensitive knowledge and whose conceptions of ownership might differ from those of the curator. For example one respondent noted that

‘[a member] has put a picture on the cover of a publication. He will be fined for that, because the artefact shows a ritual/secret process. *This is despite the fact that it was a three-dimensional artefact sold to him.*’

As another noted

‘How do you know this is sensitive knowledge? – the problem is that you might not know until it gets public [...] People think that access is to take everything you have and put it online. If we did that, we would alienate many of the communities we deal with [...] For instance, we have a photo in the Simpson Collection showing relationships among people that these same people deny having had. They see in these pictures the political rather than the daily practices [...] Also, the museum is thought to be a neutral place for many communities because it’s so far from them. They sometimes prefer their objects to be there rather than on the island of their neighbours. Then to put everything online would be the best way to ensure not working with them anymore’.

As far as the respondents were aware, no conflict ever arose from CurationProject having publicized specific objects but their ‘ethical obligation to respect creators and communities of origin’ nevertheless constitutes a barrier to the circulation of content over e-media and was currently limiting their ambitions in this direction. CurationProject thus acknowledges being caught between two trends: one driven by new technologies encouraging the widest distribution; the other advocating a case-by-case approach based on informed consent and the respect of ownership claims.

In similar vein a member of AnthroProject commented on how informed consent might not necessarily be a solution. For him ‘informed consent’ is meaningless because the person who records and makes data public over the web cannot foresee its use. Indeed whilst there are strict definitions of informed consent and in the UK, he commented that consent has to allow for secondary analysis if the use is envisaged; one has to specify what data are useful for what purposes and this is often impossible to predict in the long term. ‘Whom to should he ask in the first place when the identity of some agents cannot be recovered? Whom to consider the descendants of producers of the past centuries?’

Although quantitative social scientists such as those running SurveyProject also derive their data from people, the issue is less acute as people are anonymized in the process, making data somewhat ‘claim-proof’. ‘The way we deal with the law that information on individuals cannot be kept is that we separate the database of respondents from the one on the results. We anticipated the change in law’, a member commented. This often proves impossible when the identity of individuals or group is *inscribed in the data*, as it is often inscribed in CurationProject’s artifacts. Corti has pointed to the same problem regarding the archiving of video or even audio materials which are almost impossible to anonymize

‘Anonymising tape recordings is vastly time-consuming and prohibitively costly. Blanking out identifying information on analogue media is also rather pointless as it distorts the data. New kinds of software are now available which enable researchers can edit, anonymise, label and copy their own digital data with far more ease. However the task is still labour-intensive’ (Corti 2000; 2004).

Is this likely to change with new e-Science tools? Most of our respondents thought not.

Documenting context: cooking with data

E-Science not only involves the (sometimes unworkable) disconnection of data from those they represent, but also from the researchers that collected them. How can data collected/constructed by one researcher be trusted or even understood by another? This requires not only visualizing data in intelligible forms, but also making explicit their context

of production and setting-up appropriate systems of assessment. It is quite important to remember that 'Datasets that are incomplete, inconsistent, or inaccurate *can still be useful* to those that are aware of these deficiencies, but can be misleading, frustrating and time consuming for those who are not' (Missier, Embury, Greenwood 2005). This issue proved key across all disciplinary contexts and although the need to make data's *provenance* explicit might be a common issue, disciplines' ability or predisposition to do so was not.

For example users of SkyProject needed to know atmospheric conditions, which detector was used, the quality of instrument (imperfections), calibration algorithms, filters used and so forth. As one scientist on the project put it

'The most basic level is how good the atmospheric conditions were (clear night, good site). [...] The second criteria of quality is the instrument. Unless you're talking of a pristine photographic plate, with electronic data, no electronic detector is perfect. One needs to fix imperfection and calibrate and one gets to know a detector and how to calibrate it.'

A developer similarly remarked

'At the user workshop, people wanted to know what algorithm had been used. It was built in the program but they say: "we need to know". They don't trust other people's calibration if they don't have the algorithm'.

Since a colour picture is itself 'constructed' out of several images created from the use of several filters, an expert will want to know which filters have been used, especially since filters have developed differently over the years. A re-user will thus want to know how much an image has been 'cooked' because they might have preferred a previous stage depending on their research questions.

The same was true of SurveyProject where trust and reliability of data are also about making explicit decisions and processes underpinning their production. Different factors explain the success of SurveyProject in this respect such as the long tradition of depositing data in quantitative social science and the fact that the survey was always intended for third party analysis. But more importantly, it is the fact that those designing the survey, those carrying it out, those processing it and those analyzing it are different, which forces each of them to make explicit in documents the ways it proceeds, even if just for the sole sake of the smooth completion of the survey. This is largely done through persons (the survey is a rare dataset to have dedicated staff to answer questions of any of its users); through events (user groups meetings and workshops on the use and usefulness of the survey); through documents: documentation and questionnaires posted online of which the survey manager said

'we're concerned about encouraging people to document their surveys properly. As part of the quality control of the survey, we document what happened. We produce things like the detailed process document, which is up on the web...quite a lot of tedious information about the details of the sampling, the weighting, response rates and all the rest of it, so that anybody coming to the survey who wants to say "oh, what about this?". They can have a look at the quality survey for the info in there.'

Finally, for those running SurveyProject, trust in data will strengthen as *good practices* and *standards* are established As the manager said

'It's part of trying to establish standards that we think people who are running these sorts of big national surveys should follow, so that things don't just get lost in the mist of time. And so that people don't go: "oh, I don't know why it's like that; the people who made the decision aren't there anymore" (...) There are a set of generally accepted practices in

terms of ways of designing questions, ways of dealing with data that you would expect any reputable survey to follow. And a lot of that has been backed up by tons of experimental research of various kinds (...) You can see the effect of doing things in different ways.'

In both cases the extent to which the data are *cooked* and by whom and with what is key to the re-usability of the data.

AnthroProject in the other hand addresses this issue from a rather different perspective. One respondent commented that one could not impose fixed procedures in anthropology: 'all you can rely on is the person to have ethics: it is that person who makes decisions and selections every day. The anthropologist is a walking filter.' By contrast, the anthropologist would thus have developed ways to auto-audit. Contrary to 'hard' sciences, in anthropology our respondents suggested that experience of the field cannot be reproduced and therefore verified. But neither are methods crucial to formalise as, contrary to both Star and Survey project where data users are concerned with demonstrations and measurements, anthropology was seen as more about explanation and interpretation. As an informant said: 'there are no "right or wrong" or clear "yes or no."' As another remarked, there is therefore a tremendous amount of trust in anthropology with nobody wanting to 'check' your data. It is a moot point as to the extent that this is true of both Star and SurveyProject.

The most contentious issue therefore remains the *context* needed to make qualitative data 'reusable' as compared to quantitative ones. For some, one doesn't need

'to have been there [...] documentation of the research process provides some degree of the context, and whilst it cannot compete with being there, field notes, letters and memos documenting the research can serve to help aid the original fieldworks experience'. For others "context" and "reflexivity" are understood to form the boundary between quantitative and qualitative data, and it is the apparent impossibility of archiving content and reflexivity which renders qualitative data problematic to reuse' (Bishop 2005).

Interestingly the UK Data Archive Qualidata service defines context by ten criteria, which generates ten mandatory fields to provide when depositing data. When it comes to eliciting context, anthropologists are good at recording the conditions of collection of their data: there are fieldwork notes, but also diaries that 'put everything else in context'. However, during an interview, members of AnthroProject explained what had initially triggered their interest for some data they recorded in the field. When asked how they could convey alongside the data the whole story that they reported, they showed a book they had published on that. Rather more than the UKDA imagined in the way of context... Anthropologists are thus good at providing context for their data but as another respondent pointed out, contextualizing data fully in anthropology would essentially come to reproduce the world. And to what extent do digital repositories intend to do this?

Conclusions

In conclusion we can see how the future of e-Social Science, and the data it chooses to make re-usable, is going to depend very heavily on the existing practices of the disciplines that encounter it.

In the introduction and throughout most of the paper we have talked as if the distinction between quantitative and qualitative approaches is relevant to their response to e-Science. On the surface this appears true. However, this divide disappears if one considers that in both cases data is not self-contained and self-explanatory knowledge that can easily be circulated,

but each time needs complementary external information to be understood or trusted. If certain domains benefit from data that meets some basic requirements of eScience (data which is born digital, highly visual and coded in a shared and somewhat standardized language), as for other domains, data's mode of production and everything that makes it as it is, will need to be made explicit if it is to be trusted. In other words, even in 'hard sciences', numbers and observed raw data aren't self-explanatory and self-legitimizing and the degree and ways in which they were constructed will need to be stated. However, one of the pitfalls might be to think that specific ways to visualize, or to describe through a specific ontology, or to contextualize with a specific set of criteria, will prevent loss of information or quality.

If data circulated through electronic media largely depends on a multiplicity of external data to contextualize them and if eTechnologies will certainly add up to existing technologies rather than replace them, then the focus of research might need to be shifted. One might not want to research how Science could be done online or how data can ultimately be visualized and described, but rather how to distribute different kinds of data across different technologies and how to redirect one technology towards another. Suggestions were made along this line. For example a member of Qualidata suggested that secondary analysis of qualitative data could be a necessary pre-cursor to or complement the collection of primary data. In the case of 'hard sciences', querying databases might help locate problems worthy of investigation or further support demonstrations. For the Social Sciences, the idea emerged that electronic media could distribute the sections of research usually rejected by traditional publishers, such as the methodological sections introducing PhD theses..

As a member of CurationProject noted, the problem with eScience might be the claims made on what it is to fulfil. eTechnologies are unlikely to be global multi-purpose tools able to condense all others, but are nonetheless valid and worthy tools to perform specific and local functions. In the field of museums, they make internal management of collections easier; they allow them to travel under digital forms, whilst actual objects rarely move. Connections with other objects are more easily made. A senior manager in SkyProject expressed the same idea that if there was disappointment regarding 'universal' solutions such as ontologies, eScience will probably allow users to run data through the pipeline so that actions previously performed on it could be undone as necessary for individual new research.

As ever scientists will turn technological innovation to their specific needs and uses. The extent to which these will prove revolutionary as opposed to evolutionary remains to be seen.

Acknowledgements

This research was supported the ESRC e-Social Science Programme funded 'Entangled Data and Knowledge' project, **RES-149-25-1002** - <http://www.essex.ac.uk/chimera/projects/edkm/>

References

- Becher, T. (1987). The disciplinary shaping of the profession. In B. R. Clark (Ed.), *The academic profession* (pp. 271-301). Berkeley: University of California Press.
- Bishop, L. (28 September 2005). *Is Secondary Analysis Second Best? A case study of reusing qualitative data*. Cresc Methods Workshop Re-using Qualitative Data, Manchester.
- Brenneis, D., Marcus G. E. (2005). In Between, and On the Margins of, the Shining Centres on the Hill. *Anthropology News*, September 2005.

- Coopmans, C. (2006). Making mammograms mobile. Suggestions for a sociology of data mobility. *Information, Communication & Society*, 9.
- Corti, L. (2000; 2004). Progress and Problems of Preserving and Providing Access to Qualitative Data for Social Research - The International Picture of an Emerging Culture. Forum: Qualitative Social Research. 1.
- Fenstad, J. E. (1995). Relationship between the social and the natural sciences. *European Review*. 3: 61-71.
- Fry, J. (2006). Coordination and control across scientific fields: implications for a differentiated e-science. Hine, C. (Ed.) *New infrastructures for knowledge production: Understanding e-science*. Hershey PA, USA, IDEA GROUP INC.
- Hine, C. (2005). Material culture and the shaping of e-science. *NCeSS International Conference on e-Social Science*.
- Jirotko, M. (2005). Organisational and technological challenges of large-scale multi-disciplinary scientific research.
- Purdam, K., Elliot, M., Smith, D. Pickles, S. (2005) Confidential Data Access, Disclosure Risk and Grid Computing, *UK e-Science All Hands Meeting 2005*.
- Latour, B. and Woolgar, S. (1979) *Laboratory Life—The Social Construction of Scientific Facts*. London: SAGE
- Missier, P., Embury S., Greenwood M. (2005). An Ontology-Base Approach to Handling Information Quality in e-Science. e-Science All-Hands Meeting 2005, Nottingham.
- Strathern, M. (2005b). Useful Knowledge. Isaiah Berlin Lecture 2005. British Academy.
- Thomas, N. (1991). *Entangled Objects: Exchange, Material Culture, and Colonialism in the Pacific*. Cambridge, Harvard University Press.
- Thompson, G. F. (2004). Getting to know the knowledge economy: ICTs, networks and Governance. *Economy and Society*. 33: 562-581.

Appendix: Field site descriptions

SkyProject – a £10M project initiated in 2001 by a consortium of 11 departments. Its distributed team composed of scientists, software developers and managers aims at building the infrastructure for a data-grid for the UK, which will form the UK contribution to a global observatory. It works closely with similar projects worldwide through the 'International Alliance'. The infrastructure developed enables the first beta-users to perform queries across distributed datasets through the SkyProject portal.

SurveyProject – This project is a resource centre producing a large scale complex survey dataset released every year through the UK Data Archive. eTechnologies, although not yet grid technologies are used all along the chain through which data are collected, processed, released, and analysed

CurationProject: an activity which has been digitizing records on its collection for more than twenty years and makes them available through an online database. With pictures now being

added to the documentation, the project is conceived as an additional 'collection'. The curator has also recently initiated two new projects by which the museum's database will be open to a community of researchers, artists and communities' representatives from around the world so that their alternative expertise can be recorded at the core of the museum. Another project aims to explore local modes of production of knowledge by distributing RSS technologies to members of a specific network, participating in and recording exchanges.

AnthroProject: here we studied an anthropologist and his students who undertook to digitise and distribute all the materials so far collected during their academic careers, including fieldwork notes. Some of the professor's students opened cultural areas based sub-sites under the umbrella of the main project. They developed a wiki, a forum, a proprietary probabilistic search engine designed in partnership with a consultancy and also participate in the Dspace worldwide digital archive, among others.