Data as Representation: Beyond Anonymity In E-Research Ethics[1]

A. Carusi
Oxford e-Research Centre
7 Keble Road
Oxford
OX1 3QG, UK
Oxford University
annamaria.carusi@oerc.ox.ac.uk

Abstract

As e-research practices begin to be adopted in medical and social science research, a number of ethical challenges are being encountered, and doubtless there are several hidden ethical implications which will emerge as e-science matures. Concerns so far have centred on anonymity, confidentiality and privacy. This article proposes a representational framework for thinking about data in e-research which takes it beyond this concern. It proposes five representational models which can be used to probe the ethical issues around data derived from human subjects, and as an ingredient in ethical decision-making in this domain.

Introduction: E-Research

There are currently initiatives in the UK, USA and Europe to extend to medical research and health care, as well as sociological and anthropological research, the tools and technologies of so called 'e-science' or e-research.[2] This shift has brought with it ethical challenges which researchers, institutions and ethicists are hard-pressed to meet. The ethical challenges are generally placed under the rubric of privacy, confidentiality and anonymity. However, underlying these is a dominant conception of data as

[1] Research for this paper was supported by ESRC grant RES-149-25-1022.
[2] See for example www.nesc.ac.uk and http://www.eu-egee.org/.

information. This article proposes an alternative conception of data as representation which widens the scope of ethical considerations to be taken account of 1) in the design of the tools and technologies for e-research that deals with human subjects, 2) in the ethical guidelines for researchers and institutions, and 3) in the general understanding of ethical aspects of research of this type and responsibilities of both researchers and research participants.

e-Research was originally thought of as science carried out by means of high performance computing which required a 'Grid'. The Grid was initially envisaged as a computational grid, that is, distributed sets of computers co-operating on a calculation. Currently, there are shifts away from a grid as such, and reliance on other Internet technologies (such as 'cloud' technologies), and the infrastructural aspects of this form of research are focused upon. Above all, e-research technologies are thoroughly social and institutional as they relate to access to research resources. As such, the way in which these relations are understood is crucial to the design of e-research tools and technologies.

The e-research applications that are the focus of this paper are those which make use of one or more of the following:

- Large-scale databases (video, audio, image, survey questionnaire responses, biographical, demographical and health information, etc.)
- Resources for data-sharing, either for data-mining, for re-use of data, or for collaborative analysis and interpretation of data.
- High performance computing or computational resources such as metrics, modelling, simulation and visualisation.

Whereas in the initial natural and physical science applications the most important social implications are to be found in the sphere of co-operation and collaboration, and the ethical implications are primarily in the domain of trust among researchers in their institutional settings, the broadening of e-science from natural science to health and social sciences has brought with it the entire gamut of social relations between researchers and others who manage and deal with data, and the human subjects from whom data are obtained. In the UK, e-social science projects are funded and promoted in various ways, with a major initiative in the UK being the National Centre for e-Social

Science, under which there are 7 large scale projects, one 12 small scale projects (see www.ncess.ac.uk for a description of these projects). It is the NCeSS projects in particular that have formed the background of this paper. Other important examples are to be found in anthropology (for example, an extensive digitisation project undertaken by the Cambridge University Museum of Anthropology and Archaeology, and several other examples on the Internet) and the quantitative data base of the British Household Panel Survey (Carlson et al., 2006). However, current digitisation drives by funding agencies who make it a condition of funding that data are archived in a data base, and by public health services which are digitising and archiving health records will result in massive amounts of information about people being grid or Internet-enabled, with the attendant issues, problems and concerns relating to anonymity, confidentiality and privacy (Carusi & Jirotka, 2007).

Among the lessons learned from projects in the medical area, such as e-Diamond, a project to build a federated database of mammograms for distributed readings and training (Jirotka et al., 2005), is that legal and ethical restrictions on what can be done with data obtained from human subjects could severely restrict e-research applications of health and social science (Hinds et al., 2007). While the only restrictions on physicists' sharing and re-use of data are their own collaborative practices, there are currently strict limitations on the sharing and re-use of data obtained from human subjects, and even where there are not, there may still be ethical reasons that preclude this. However, the effect of legal and ethical restrictions and personal ethical reservations of some researchers will not be to shut down health and social science applications of e-science, but rather to shift them towards methodologies which are more amenable to these cyber-infrastructural environments.

We should expect to see an increasing mix of quantitative and qualitative methodologies, with images, audio and video which are often used for qualitative social science research, being subjected to increasing quantitative measurement. Data merging and mining are being increasingly used, as are other computational resources such as webmetrics for social network analysis. There are extremely important shifts in the mode of doing health and social science unfolding at the present time, and with them will come extremely important social, ethical and political shifts too. It is important to keep as rich

a variety as possible of different types of research and different methodologies alive and viable on the Internet and on specialised research grids. Important, that is, both for a rich and nuanced understanding of human beings and societies and for the social and political function of those understandings. In light of this, it may be necessary to re-think some of the ethical principles that have informed research practices, or to re-think the relationship between researchers and subjects of research.

These developments are currently unfolding and as yet cannot be studied with the cool detachment of hindsight. Instead, some foresight is needed to try to project what hidden implications these tools and technologies may have in order to take these into account for both the design and development of the technologies and the way they are socially and institutionally embedded. Indeed, this has already been sharply felt in Internet research ethics. In order to try to address these challenges, the Association of Internet Researchers set up a committee, led by Charles Ess, to offer guidance to researchers in facing a very different ethical landscape (Ess et al., 2002). E-research differs in important respects from Internet research, as it does not (necessarily) carry out research on subjects already on the Internet with the purpose of understanding Internet behaviour. Social science use of e-research technologies is different in that the research is mostly carried out on subjects in the offline world; it does not study subjects on the Internet, but uses cyber-infrastructural resources for the archiving, analysis or interpretation of data. However, there is a need for a similar process of consultation and ongoing debate in order to try to meet the ethical challenges attendant on the new technologies. The aim of this article is to broaden the range of factors that are traditionally considered to fall under the rubric of research ethics, beyond the central three issues of anonymity, confidentiality and privacy, to more general representational aspects of data.

The article begins with a consideration of notions of identity in research relating to human subjects and defends the ethical significance of thick identity. In the second section, thick identity is in turn related to a view of data as a form of representation as well as only information, via the notion of 'representation as'. In the third section, I consider the possible effects of digitisation on data, the ways in which it can be processed which are relevant to an understanding of data as representation. In the fourth section, an

outline of five representational models is given as frameworks for understanding different forms of data and the relation they have to researchers and research subjects who are implicated in them. A wider study would consider institutional aspects as well, but this cannot be covered here.

Identity

While not in any way diminishing the importance of the problems of anonymity, confidentiality and privacy, this paper suggests that there are ethical issues and concerns beyond these, which emerge when data are regarded as representation rather than as information. Concerns around anonymity, confidentiality and privacy revolve around the possibility of re-identifying individuals, and the information that can be obtained about re-identifiable individuals. I distinguish between two senses of identity: thin and thick. The distinction can be seen from the point of view of what is believed or known on the basis of data. Thin identity is the identity of a particular individual as a re-identifiable entity. Proper names pick out particular individuals and have to do with thin identity. Thick identity is a matter of that individual's experience of their own personhood, their own subjective or psychological sense of who they are.

I wish to suggest that thick identity can be a matter of ethical concern even when it has been detached from thin identity. Thus even if anonymity is guaranteed, there are ways of treating thick identity that it would be appropriate to make an ethical judgment about. One aspect of thick identity is the narratives that people use to make sense of their lives and circumstances. Taking over or appropriating images, quotations, stories or other ways in which thick identity is mediated, possibly for some purpose to which consent has not been given, is an instance of a use of anonymised data which could be considered as apt for ethical judgment.[3] Of course, we are accustomed to this happening, for example in journalism and in biographies, but firstly, the fact that we are does not mean that they can therefore no longer be considered as cases for ethical judgment; and secondly, it is important to recall that we are dealing with research cases, and the type of agreement or contract that there exists between researchers and their subjects is not the same as that

[3] See Tomaselli (1996) for an interesting treatment of this topic. For narrative discourse and identity, see Bamberg (2004), and for an online development of this view, and the ethical guidelines given for it, see www.talkinglongterm.co.uk.

which exist between journalists and their interviewees. Considering this kind of appropriation as something of moral relevance has to do with the representational content of the data, rather than with information relating to an individual.

Another way in which thick identity is experienced is by means of identification with a group, or a collective identity. Our current social and political climate means that head scarves are as much a matter of the individual identity of the women who wear them, and of their group identity. The head scarf is a metonymic signifier of something that goes beyond allegiance to the group, but of an experience of personhood in terms of the group. The ethical questions over how groups are portrayed are now familiar from decades of cultural studies into gender, race, and homosexuality. Thus, even if no particular individual is identified by some way of handling data, some aspect of their thick identity – as carried by data of some form – can be used to construct an identity for a group and this construction has ethical implications.

Information and Representation

Data are normally seen as information rather than representation. However, thick identity is often a matter of representation rather than information. For example, the sense of self is often expressed and formed in a kind of personal narrative, a form of representation with great psychological power. Images and metaphors are important in this too. The distinction can be put as follows: the thin identity of an individual has to do with the relation between that particular individual and the fact that they have a certain medical condition; the thick identity of that individual has to do with his or her representation as a victim of that condition, rather than as a fighter, or survivor, or simply neutrally as having it. This is where the notion of 'representation as' as it is theorised in philosophy of art and cultural and critical studies has potential to be useful as a way of understanding fundamental ethical concerns. 'Representation as' derives from the insight that there is not a single or unique way to represent an item or a person (Goodman, 1976; Gombrich, 1983). Given the multiplicity of ways in which people can be represented, the fact that they are represented in one way rather than another can be motivated by values, interests, prejudices which are not obvious either to those being represented nor to those doing the representing. Verbal and visual representations are a crucial arena for the formation and

playing out of personal and social values. The social values embedded in representations are often the hidden and implicit motivators of explicit moral judgments. Philosophers have different ways of distinguishing different aspects of the moral terrain, but one way to see this is in terms of Habermas's conception (drawn from phenomenology and hermeneutics) of the lived world as the context and horizon of expectations in which group solidarities are formed around certain values as a background for ethics proper (Habermas, 1988).

It is by considering data as representation that some of the ethical concerns around it that would otherwise remain hidden come to light. This attunement to the ethical dimension of data as representation has an affinity with disclosive ethics with respect to technologies (eg Brey, 2000): just as hidden and implicit values are embedded in the design of technologies, so are they embedded in data considered as a form of representation.

In this section I have argued that thick identity is closely related to 'representation as', that is, to the way in which subjects are represented, and that representation is value-laden in a way that makes it apt for ethical consideration. I will not try to broach the question in this article of what kind of ethical consideration or ethical action follows from this view of data as representation. The point I make is a small one: even if anonymised and not traceable to any particular individual, the representational character of data has ethical implications which need to be considered as we try to work out our overall reactions to e-research in the social and health sciences.

Possible effects on data of Digitalisation

Digitalisation processes allow for a myriad different possibilities of manipulation of data, from the very basic construction of data bases, to more intrusive manipulations (Gross, Katz, & Ruby, 2003). In this section I consider standardisation and reification, re-contextualisation, manipulation, and computationally assisted interpretation. This is not intended as an exhaustive list, though they do seem to be central to many e-science informed modes of research.

Standardisation and reification

In the case of e-social science, a particular set of issues emanate from the needs of standardisation. A shared database is not useful unless there are standardised ways of presenting and dealing with data. If data are to be made available in data bases for possible re-use or collaborative use, they must be standardized.[4] In addition, the computations that are accessed via the grid could require data that are delivered in a certain way: for example, in the case of a multi-modal analysis of natural language case, a computation to measure the range of the head nods that accompany language requires that the image be taken in a particular frame, from a particular distance and angle (Knight et al., 2006). How it is standardised makes an enormous difference. The choice to present a three dimensional object in two dimensional format; the choice to have all images of head nods framed in a particular way, the choice to cut and crop in a particular way, to zoom in on some aspect rather than on another, the choice to associate particular labels / metadata with images: these are all examples of digitalising processes.

Standardisation is connected with reification. Reification occurs when the aspect under which something is known comes to be identified with that thing. For example, the anthropologist Marilyn Strathern, describes it as follows:

> By reification I simply intend to point to the manner in which entities are made into objects when they are seen to assume a particular form ('gift', 'exchange'). This form in turn indicates the properties by which they are known and, in being rendered knowable or graspable through such properties, entities appear as "things." (Strathern, 1999, p. 13)

Standardisation feeds into reification because it forces all items in a data base to be represented under the same aspect which then defines by which properties the items represented are known. Reification of the subjects represented under those aspects is facilitated. Reification can sometimes result in data acquiring an independent status from subjects from whom it was derived. This has different ethical consequences depending on

[4] In addition, it is often required that the way in which data are standardised be transparent to other researchers. The provenance and history of data are extremely important for other researchers if they are to trust it. See Carlson & Anderson, 2006, on astrophysics, and Jirotka et al, 2005, on mammographies.

the type of data they are: for example, a discrete item of data (image, video, quotation) embedded in a larger data base, or a series of data all connected to the same person.

Re-contextualisation

De-contextualising: e-research data are made accessible to researchers beyond the initial data gatherers. They are shared either for collaborative research, or for re-use for other research.

Re-use of data means taking them out of their context of collection – that is, out of the situatedness from which they emerged. Since context plays a significant role in specifying the representational content of items, this would evidently have an impact on the content, meaning and significance of data. Using audio or text alongside images cannot but suggest that they are meant to illuminate each other, and should be interpreted accordingly. Any film maker knows that; however, the new digital media make these possibilities available to anyone with the technological means and also make it possible to take apart contextualising features of a representation. De- and re-contextualisation can also have ethical implications. Contextualising a photograph of a slender girl in a data base of sufferers from anorexia cannot but represent her as an anorexic.

It is instructive to use anthropology as an example, as museum collections have historically been less regulated than other forms of social science, and can be instructive as to what can happen in the absence of regulation. A random example is the Smithsonian Natural History Museum's online exhibition dedicated to 'African Voices', where various images of people and artefacts, as well as voices, can be accessed.[5] The site presents an interesting case for political and ethical analysis of the effects of digitalisation, where the hypermediated contextualisation offers itself to critique in terms of its implicit values regarding the representation of 'others'. The point here, however, is that digitalisation makes possible endless contextualisations and re-contextualisations, with not only the site presenting its own array of contextual possibilities but also making it possible for viewers to continue the process of de- and re-contextualisation. For

[5] http://www.mnh.si.edu/africanvoices/

example, it is possible to download any of the images on the site – in fact, sometimes the viewer is invited to do so: the image of a mudcloth, for example, can be downloaded as a screen saver. This is quite a radical de-contextualisation of these images, one which results in their representational content being quite drastically affected (one can't see it is a cloth at all, let alone a mudcloth; the shapes on it play out their significance among the shapes of the viewers life, and not those of the producer's (cf Geertz, 1976). What can be done with the relatively insignificant mudcloth can be done with any of the images or audio. The ethics of appropriating representations of others and of their world are debatable and I am not suggesting that the ethical implications of the new possibilities of interacting with data brought into being by digitalisation are necessarily negative and exploitative. My point here is that the digital medium makes these re-contextualisations possible, with the concomitant effects on the meaning of the data, and the associated ethical issues.

Manipulation

Standardisation and (re-)contextualisation in data bases are both facilitated by the greater possibilities for manipulation of digital data. This is particularly evident in images: these can be cropped, the colour can be changed, montage is far easier, and so on (Gross et al., 2003). It is in view of the possibilities of manipulation that it is important for researchers to have a full biography of data that they use or re-use in their own studies. It is not scientific fraud that is the main ethical worry here, but rather the type of manipulation that falls within the bounds of normal scientific procedure. Any of these manipulations for the purposes of standardisation could result in differences in the 'representation as' aspect of data.

Interpretation and computation

e-Research applications use digitalised data in order to be able to use on them analytical and interpretational resources that would otherwise be inaccessible or difficult. In the first place, collaborative interpretation is made possible by the e-science infrastructure. For example, this was a feature of e-Diamond, the grid-enabled infrastructure to allow mammographies to be read and interpreted at different sites of expertise. A standard kind

of interpretational resource is annotation; in e-science applications there are two possibilities that stand out: collaborative annotation and multimedia annotation (for example, running transcript and video or audio side by side and being able to annotate both at the same time). Annotations then become a part of the context of the data, and inflect its representational content. With these two possibilities come other ways in which 'representation as' can be affected: first of all by the intervention of other researchers besides those who gathered the data from subjects and have a direct relationship with them; second by the juxtaposition of different media.

Data-mining is another technique which is facilitated by digitalised data in large data bases. Data-mining can latch on to the annotations or other tags attached to data by researchers. Data-mining leads to re-contextualisation, and sometimes it can lead to a re-contextualisation which results in the data coming to be categorised under social groupings which the data subjects – nor indeed the researchers who dealt with them directly – could foresee (Tavani, 2004). The representational content of the data thus is inflected by a social group which is not consciously part of the data subject's thick identity. This does not mean that it not a part of thick identity: most people of the petit bourgeois class in Karl Marx's time would not have recognised themselves in the class structure as described by Marx, but this does not in and of itself invalidate what Marx had to say about their consciousness. But this in itself shows that these re-groupings are not value neutral.

Standardisation, reification, re-contextualisation, manipulation, data-mining, annotation all have both methodological and ethical implications. They can privilege or occlude some analytical and interpretational possibilities in specific studies, and lead to systematic overall analytical and interpretational trends in social science.

The fact that these technologies are not ethically neutral need not lead to dismay and attempts to curtail them. They could have positive as well as negative implications. For example, digitalisation can break previous norms of standardisation (something which art has always done) and disclose what these norms are, thereby de-mythologising them; the re-configurability of data collection means that different aspects and stories about the people represented can emerge, and the manipulability of digital images and

other information means that they can also be re-appropriated by their subjects. There is also a greater interactivity with data and data bases, and with digital collections, access allowing. These are powerful media, and because they are, the question who has access to them, and what sort of access (viewing, editing, downloading, etc.) is paramount.

This brief overview of some of the possibilities of handling data opened up by digitalising and Internet-enabling them indicates that there are ethical responsibilities around the way in which they are handled which go beyond ensuring anonymity, confidentiality and privacy. It is still a question towards whom (or what) these responsibilities are to be directed. This is not divorced from the question how the representational content of data is produced or constructed, as I go on to show in the next section.

Representational models

It is often the case that people who are in ethical decision-making positions – including designers and developers, researchers and their subjects, and the ethics committee members who must vet their research – assume certain models, or have certain implicit expectations of how representations work. Because the model or models they have in mind are not explicit, it is not clear to whom or what ethical consideration is due. To highlight the fact that data can be representations as well as simply bare information, while at the same time claiming – as I have – that this is an aspect that needs to be kept in the foreground while considering how or whether data should be digitally enabled and archived only goes a small way to broadening the scope of ethical negotiations around data. Representation is not representation *tout court*, as art history, literary theory and philosophical aesthetics attest. In the case of social science data, there could be competing assumptions and preconceptions about how representations function among different stakeholders. Different people will have a different implicit model about how representations operate, and this will be part and parcel of their ethical attitude towards data considered as a form of representation.

In this section, I describe five representational models, which are broad frameworks for the conceptualisation of data as representation, focusing in particular on the relation between representer/researchers and represented/data subjects. The models

are offered as heuristic tools for thinking about the link between representational and ethical considerations. It is not, however, the case that data sets will fall uniquely into one or other model; rather it is more usual that there will be a difference between the model that different stakeholders implicitly apply to the same dataset, and there may also be tensions and competition between them. The application of specific representational models can be affected by a number of factors, not least of which are the power relations between those in the representational context.

The representational relation defined by who represents whom is fundamental in determining representational content; among other things it also tells us much about the context of the representation, the expectations with which it was made, and the purposes. For example, a family snapshot has a different representational content to a passport photograph even if it looks the same. The representer/represented relation can be aligned with relations of power and the playing out of values in representations is often a matter of the distribution of power across the research environment – it is, after all, for that very reason that human subjects are protected by ethical regulations. This does not mean that power relations are always in favour of researchers; for example particular social groupings of potential data subjects can exert a great deal of influence in having some research but not others pursued (for example, research into so-called 'lifestyle' illnesses). In addition, researchers do not do research in a vacuum but are dependent upon research communities and institutions of different types.

The five models I consider are:

- Naturalism
- Isomorphism
- Figuralism
- Constructionism
- Interactionism

Naturalism

Naturalistic representation could also be labeled 'the disappearance theory of representation' as it prevails when representations do not seem to be representations, or

are so closely connected to what they represent that they seem as faithful as a representation can be. In these cases, the representations seems to be more like a screen or window and the surface of the representation seems to be a neutral 'giving over' to the thing represented, something that we simply look through in order to see the 'real' subject of representation. Naturalistic representation draws on resemblance theories of the way in which images and other visual media gain their representational content, and / or by the naturalistic variant thereof, which sees continuity between the way in which objects are perceived in the actual world and in art and other visualisations (eg a realistic picture of a face) (see for example Hyman 1997; Peacocke 1987).

In the case of photography and film, there is the further dimension of the causal connection that there exists between the colour, lighting and structural features (such as spatial arrangement) and the real world state of affairs of which they are the photograph or film. This creates the effect of a kind of immediacy in photograph and film: a capturing of something as it was at a particular time: the content of the photograph or film seems to be 'this-here-now'. The 'this-ness' and the '(t)here-ness' seem to be inscribed into these media in our experience of them (Barthes, 1981). This can also occur in recordings of voice, which can have an effect of presence in representation.

On a naturalistic view of representation, features of the subject are mapped onto features of the data. A face is recognisable in a naturalistic representation, particularly if it is a photograph or film. The data are also taken as reflecting aspects of subjects' persona: for example, their mode of dress, their gestures or other aspects of behaviour, and thus also to be revealing about 'thick identity'. It is part of the naturalistic model of representation that representational content is determined by the thing or object represented rather than by the representer, who – together with the technologies used to create and to present the representation – recedes into the background, or indeed does not figure at all.

Naturalism has many limitations as a theory of representation, since representations are not as neutral or transparent as naturalism suggests, and the representer as well as technologies of representation are an essential aspect of the way in which things get represented; *however it is the implicit 'theory' of representation which*

is most often spontaneously adopted by research subjects. Because of this, it cannot be discounted on the basis of a rival theory of representation. Indeed, because it is so deeply embedded in the phenomenology of representations, in which it is ordinarily not salient that representations are constructed in various ways, the concerns about the way in which subjects are represented should be taken very seriously indeed. On this model, because the representer's role is not phenomenologically salient, it can be held – in particular by subjects whose features can be immediately recognised in the data on the basis of resemblance – that the represented subject has the primary stake in the representational content, and ethical concerns are directed towards the subject of representation, with respect to both forms of identity, thin and thick. This is particularly the case because any 'representation as' content will appear naturalistically, and so as simply the way the subject is. If the sensitivities of a stakeholder who holds an implicit or explicit naturalistic model of representation are respected and acted upon, it will be important that discrete data should not be isolated from subjects, and the relationship between subject and data needs to be kept live. This would preclude re-use of data by researchers who have had no direct contact with the data subjects; it also precludes images or other representations being placed in data bases not agreed to by the subjects, as this will affect the way in which they are represented (even if they cannot be re-identified).

Isomorphism

This is an interesting view of the representational relation between patients and their digital medical records put forward by Eike-Henner Kluge (2001), and inspired by Wittgenstein's picture theory of meaning.

Kluge claims that since (1) digital patient records bring together items that in paper-based records would have remained separate; (2) they allow for a spatio-temporally 'joined-up' (in space and time) 'view' of the person qua patient, (3) combinations and manipulations are holistic (that is, a manipulation on one part affects the whole), digital health records are not simply copies of the features of the patient's health status, but actually share some features – such as the combinatorial features of representational elements) with the elements represented. This makes electronic patient analogues of patients. The analogical relation is limited to 'those aspects of the patient which have

generated data in the health care professional / patient encounter (Kluge, 2001, p. 30) and does not include the patient as a person.

This goes beyond naturalism as a theory of representation, because naturalism depends on a resemblance relation between discrete items (the subject's face and her image, for example). The analogical relation is instead a many-to-many relation, as it ranges over the diverse items in a digital patient record (blood tests, X-rays, diagnoses, etc.) and the diverse aspects of the patient to which they are connected. In particular, the analogical relation holds between the relations among the diverse items in the digital health record as a whole, and the relations among the diverse aspects of the patient. The analogical relation is thus a relation between two sets of relation. There is an analogical relation between these two sets of relation because they are isomorphic: that is, not just similar but the same with respect to their form or structure. A picture of a rectangle is isomorphic with a rectangular box because it has the same relation between (some of) its parts as the box.

According to Kluge, the ethical implications of isomorphism between digital health records and patients is that the first is an analogue of the second, and that they should be treated 'according to ethical principles that are analogues of the very same principles that normally should govern our conduct towards persons in the real world' (Kluge 2001, p. 39), and not merely heuristic codes of conduct. For example, it implies, according to Kluge, the ethical principle that 'The electronic patient record should be treated never as a mere thing but always as a person-analogue in information and decision-space.' (Kluge, 2001, p. 56).

Even stronger than naturalism, isomorphism sees data as being deserving of treatment analogous with that of the subject. We have already encountered the notion of reification in the context of some of the consequences of digitalisation. Reification is a feature of digital health records, according to Kluge:

> Once it has been generated, the existence and functioning of this patient analogue is independent of the patient and functions independently as basis for interventions and decisions-making. Therefore, to all intents and purposes, it has acquired what amounts to a functionally independent status. In this sense, we are

> beginning to see the ontological reification of the patient record (Kluge, 2001, p. 33).

Thus, the digital patient record is itself a direct object of ethical concern, as is shown for instance by Kluge's Principle of Security:

> Data that have been legitimately collected about a person should be protected by all reasonable and appropriate measures against loss, degradation, unauthorized destruction, access, use, manipulation, modification or communication (IMIA Code of Ethics for Health Information Professionals).

To treat a set of data relating to a person as analogous to subjects is to treat it *as though* it is the subject. The data set contained in a digital patient record is the direct object of ethical concern, and the subject is the indirect object of ethical concern. This in no way implies that the subject is secondary to the data[6], but rather that the digital patient record mediates the ethical concern towards the patient.

This view of digital patient records has met with criticism. For example, one reviewer has criticised this account of digital patient records:

> Such a view of representation – apparently inherent to what Kluge refers to as 'an information-theoretic standpoint' – denies all the (error prone) choices, dilemma's, work, translations and (human) agency involved in the construction of records, and precludes appreciation of the well-documented need for a more empirically informed socio-technical understanding of the role, nature, and function of medical records in actual health care delivery (Van der Ploeg, 2003, p.66).

Isomorphism, thus, will be in tension with constructivism, in particular with ontological constructivism according to which that which is represented (the person, thing, event, etc.) is also a construct of research. This is just one of the tensions that may exist between competing representational models (a point to which I return).

[6] A further analogy, this time with theory of perception may help. Indirect realists do not hold that the real object indirectly perceived is secondary to the mental object directly perceived.

Isomorphism is a useful representational model because it allows a grasp of a particular feature of the Internet or Internet-related technologies, such as the Grid. Discrete items of data may not constitute items for either positive or negative ethical concern, but when several discrete items are connected up, they allow a picture of subjects to emerge which may well be an object of ethical concern, and will certainly have rich 'representation as' content (Nissenbaum, 1988; Nissenbaum 2004). The recent release by AOL of users search logs is an example: a single search is not ethically significant, but a series of searches is.[7] In the AOL case, Internet users were identified in some cases, but even in the absence of positive identification of particular subjects, on the isomorphic representational model there is reason to hold that the logs constitute an analogue of an actual person, and are therefore, as a body of data, a direct object of ethical concern, even if we do not know who the person is. We may owe these data, with the picture of a person that they present, the same kind of ethical concern that we owe to strangers, that is, to at least not harm, and in some cases try to protect.

Figuralism

A familiar topic in cultural studies, the rhetorical aspects of representation can be an unacknowledged force in the formation of representational content in the social sciences. However, the anthropologist Marilyn Strathern alludes to different figural constructions at play in the representational relation between the 'whole person' and an artefact:

> On the one hand, all over the world we find systems […] in which numerous distinct and specific items are drawn together, with the sense that nothing less than total enumeration will do. On the other hand, people can use the specifics as such to summon a larger vision of themselves. This idea is a different kind of access to totality: One artefact could be enough to point to the whole. […] Here one found another bifurcation, between wholeness imagined as the sum of parts and wholeness summoned in an individual item. (Strathern, 2004, p. 7)

[7] See for example M. Barbaro and T. Zeller, A face is exposed for AOL user 4417749, New York Times, Aug 9 2006, http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=1&adxnnl=1&oref=slogin&adxnnlx=1155744807-DKt5sZmwY2uRl/mAsmGK9w

These are the rhetorical figures of synecdoche (a figure by which a more comprehensive term is used for a less comprehensive or *vice versa*; as whole for part or part for whole, genus for species or species for genus, etc. *OED*), for example, when the group or class to which a person belongs is used to denote the person, or the other way around; or metonymy (figure produced by substituting for a word or phrase denoting an object, action, institution, etc., a word or phrase denoting a property or something associated with it; an instance of this. *OED*), for example, when a feature associated with some item or person or group (religiosity, liberalism, money) stands for the item or person or group. Anthropological artefacts often function figurally in order to evoke or express aspects of the people from whom they emanate. A chalice is an anthropological artefact which expresses a great deal about Christian communities via the figure of synecdoche. Any number of other figural relations could play a part in determining the representational content of data; for example, a brain scan can be seen by a patient as a metaphor for her personality, and indeed artists, such as Susan Adlworth, have constructed self-representations out of them. They are significant relations in that they underlie the reasons why there is often an affective connection with data, and they are often powerful just because they are not obvious. They are sometimes a reaction or counter-reaction to reification, and they are part of the ways in which subjects or even researchers may include data in their 'stories' or their overall self-representation.

Figural representational content occurs alongside or as an overdetermination of representational content of data used for scientific purposes. Figures are a connotative rather than a denotative representational device, and they are an integral part of any symbolic system, linguistic or computational. Internet presentation has its own particular rhetoric, and figural content can be affected by the existence and placement of hyperlinks as well as by the possibilities of interaction that it allows for (Burbules, 1998). Figural representational content is often an essential part of 'representation as', and is not to be discounted on the grounds that it is subjective. Apart from the fact that this often rests on a false distinction between subjective and objective, relations of trust between researchers and subjects (either as individuals or as groups – as is borne out by anthropological gathering and use of data) may depend on understanding what data means for subjects, as well as what it could mean for others. In fact, this is an important methodological

consideration in the interpretation and presentation of some kinds of data, in particular qualitative data. Having said this, the ethical concerns that figural aspects of representations evoke can be very difficult to pin down, but in the context of so-called 'digital repatriation' and digital anthropology generally, politico-ethical consideration has in several cases left it to be decided by the provenance or origin of the data. In this case figural representational content – that is, metonymic or synecdochic figures as described above – may be an indication that the provenance of the data – the group or person from whom it was acquired – carries much and possibly over-riding ethical weight (Srinivasan, 2006; Carlson & Anderson, 2007). This is not only because of questions of ownership – though this may be the case – but also because it is the thick identity of the source of the data, that is at stake, and this identity is figurally indicated in the representation.

Constructionism

Data as construct is the oppositional term to data as naturalistic / isomorphic representation. In both of these, it appears that a representation is as it is because the subject it represents is as he/she is. A photograph shows a person with blue eyes because the person of whom it is a photograph has blue eyes. A constructionist representational model instead emphasises the extent to which representations are not neutral vehicles for grasping and conveying the features of the things they represent, but artificially assemble and produce representations. A huge research topic in its own right, constructionism in art and literature is nothing new and has even become the mainstream theory of representation.

In Internet research (of people on the Internet) constructionism is motivated in part by the diverse kinds of identities that are played out in the Internet, where there is a great deal of scope for people not being what they appear, and thus for any kind of naturalism or isomorphism to be extremely problematic (White, 2002, p. 249. And explorations of identity (in my sense of 'thick' identity) endorse the idea that subjects' identities are never simply clear-cut pre-existing entities, but are generated in an ongoing bricolage or construction. This is particularly apt on the Internet.

In Internet research, taking data to be constructs results in ethical concern shifting from the subjects who generated the representations to the representations themselves, for

example, showing them to be a certain kind of political, institutional, ethical entity by doing a discursive or political critique which shows how they have been constructed. This kind of reading can be critical of the subjects themselves or of the groups of which they are representative (White, 2002).

In e-research the constructs whereby subjects are represented are not generated by the subjects themselves, and thus raise a whole series of other issues. It is simplistic to say that they are generated by the researchers individually; rather they are a matter of the whole research context. And in e-research it is essential to take into consideration the role of the technology in constructing the data. Statistical research is a very good example of highly constructed research, since it takes data, processes it, analyses it into different aspects or measurements, and re-assembles or re-synthesises these measurements. There is nothing to say that constructs cannot be highly informative. E-research is likely to encourage constructed data, even from the very basic level of allowing for greater resources for annotating and analysing qualitative data. The result of these techniques are data which have naturalistic elements, but which as a whole, together with the annotations, codings, etc, are constructed data. The manipulation, standardisation, collection, and arrangement of data that digital and Internet technologies make possible will certainly reinforce the conception of data as construct – i.e. as not deriving its content from its relation to the subjects but from its relation to researchers and their context.

In e-research this model results in a shifting of the primary ethical concern from the research subjects to the data, conceived as constructs, and the primary stakeholders in that data become the researchers rather than the subjects. The acknowledgement of the data subjects consists in seeing them as the extrinsic occasion for the data rather than as intrinsically related to them, so 'representation as' does not actually reflect on the subject (even though it may seem to, in virtue of naturalistic content).

This is the attitude that allows for much actual practice in current social science and other research, where images and other information are regularly used by researchers in publications and presentations, and treated as their own property. The constructionist model can (and often does) live side by side with other models. For example, even

though on this model, data are constructed, there may also be aspects of naturalistic content or a perception that there is by data subjects. Researchers may have a sense of responsibility towards data and data subjects in virtue of naturalistic content, but these are seen as a matter of their own personal ethics (or 'Good Samaritan' ethics (Ess, 2002)) and not something which should be subject to external ethical protocols.

Unlike the isomorphic model which implies that data should be treated as an analogue of the subject, the constructionist model has no reason for doing so as this intrinsic link with the subject is broken. Rather the data are a constructed object of the researcher (and their research context), and constructions are not the same kinds of entities as analogues. Constructions imply relations of ownership, material and intellectual, rather than relations of identity, and indeed seeing the representational content of data as constructed by the researcher brings this closer in line with legal interpretations of the ownership of data and copyright, wherein the processes of creating data and data collections are taken into consideration in deciding ownership (D'Agostino et al., 2006; Burk, 2007). Thus the sensitivities around data informed by this model would indicate that the data themselves are a direct object of ethical concern, but insofar as they are an object of ownership by the researcher; researcher in turn mediates ethical concerns towards the research subject.

Thus on this construal, obstacles to sharing data for collaborative research or re-use will have more to do with the research interests of the researcher, rather than with data subjects. Constructionism is a view of data as representation which takes onto itself the burdens of ethical qualms associated with data sharing, and reification, re-contextualisation, manipulation and the overall way in which subjects' identity is represented, and much is a matter of researchers' ethical sensitivities. This is the underlying model held by emphasis placed on researchers' negotiating archiving of data on behalf of data subjects.

Interactionalism

On an interactional account of the representational content of data it is a mistake to see data as representing the subject(s), or as being a representation of the subject(s) by researchers. Data are generated in the interactions between subjects and researchers,

within particular contexts and settings. In addition, on this view data are not only *of* someone, but always addressed *to* someone. The relation to the addressee specifies much about the data. For example Heath & Luff (2000) show that the features of medical records (what is recorded, what is left out, how it is recorded) are motivated by the fact that they are addressed to other professionals. This is in line with dialogical theories of language, such as that of Bakhtin and followers, which highlight the fact that the way in which utterances are directed towards their addressees shape their form and meaning. With this conception of data, the technologies involved in generating the data could be very important in defining the interactions around it, because technologies have specific sets of interactions built into them. Thus the possibility of collaborative interpretation of data on the grid (for example using MiMeG) will make a difference to how it is gathered, and to interactions with the data subjects, and this possibility is thus always a part thereof.

The representational content of data cannot be understood at the level of individual subjects or researchers, but rather as arising through ongoing interactions among members of a group within a context. The context ranges over immediate concrete context (the physical features of the interview situation for example), to broader contexts (social and institutional). Importantly, the content of data is seen as emergent in that it arises in the interaction (Goguen, 1997). Another important feature of this model of representation is that it incorporates the values which people hold as an integral aspect of the data that emerges out of the research interaction. For example, in ethnomethodological research which is one expression of the interactionist model, it is stressed that communication and information have an 'inalienable ethical dimension' (Goguen, 1997, p. 47), and are not abstract ideas or ideals existing in a realm separate from data and research. This means that values are embedded within the representational content of data. On this account, whether data could be shared by being collaboratively analysed would depend on the ethos of the interaction. If trust is central to the interaction, the data may be shared with others who share in the trust relation (perhaps other researchers who have the data collectors trust). However, this conception of data also means that it would be methodologically unsound to de-contextualise data from the

interactions and contexts in which it was generated. Thus the possibility of re-use in contexts entirely dislocated from the primary context is not great.

On this model, the data are the direct object of ethical concern on the part of both the researcher and the subject. The researcher cannot arrogate to him-or herself the role of mediating the ethical concerns of the subject. The subject as well as the researcher has responsibilities towards the data. There are different degrees to which this model can be put into practice. It is often used together with constructionism, but it also lends itself to a participative model of research. For example, the 'Enabling Diversity: Extending Collections Information with Arctic Communities' project under the auspices of the Cambridge University Museum of Anthropology and Archaeology is using the greater interactive capacities of the Internet, and in particular online social computing (Web 2.0) to allow user communities to interact directly with the collections, and specifically 'to add information and develop historical narratives (Corti, 2000).

The interactional model allows data subjects to be active participants in their self-representations, either as individuals or as members of a community. This exploits the full potential of digital and Internet technologies to act as facilitators of reflection on these representations and the way in which they function in the formation of identity.

This is no doubt not a complete list of possible representational models for e-research data. Missing from this picture of representational relations is the role of technologies involved, and their designers and developers, and undoubtedly ongoing study will show these to have a very significant role. However, it should be clear from the discussion that the primary objects of ethical concern shift according to the implicit representational model held by different role players: researchers, subjects and archivists.

The models may co-exist, but not always peaceably: for example, there is likely to be tension between naturalism and isomorphism on the one hand and constructionism and interactionism on the other. Differences between implicit representational models held by different role players could account for some of the struggles around data and the different attitudes towards the different tools and technologies for e-research in the social science domain. A predominantly constructivist model of data may find data archiving and data mining less problematic than a naturalistic or interactional model. For example,

data which have a predominantly interactional representational form is less likely to be apt for data archiving, and this particular technology may be resisted by researchers and participants alike. An option would be to exploit interactive Internet technologies in order to extend this to archived data.

These representational models have been outlined in a more-or-less neutral way, in that no attempt has been made to weigh one up against the other, epistemologically or methodologically. The models outlined (and there may be others) are often implicitly held within domains where decisions and choices are made concerning ethical aspects of data, and need to be 'teased out' of the domain in order to understand how these choices and decisions are being framed by all participants, be they researchers, users, developers, funders, or ethics committee members. A further role for attempting to understand which representational models are at play is for necessary conversations to be had about ethical aspects of Internet-enabled research involving data about human subjects. The models are thus but heuristic tools to probe attitudes towards representation, bring them to the surface, and to tackle the ethical aspects of data as a form of representation. It is important that the conversation not end at the point of ensuring anonymity and confidentiality.

Conclusion

This article has presented a case for considering data as representation and not only as information as we attempt to grapple with the ethical implications of e-research. A distinction was drawn between thin and thick identity, and it was suggested that thin identity – the possibility of re-identifying an individual – leads to ethical considerations around anonymity, confidentiality and privacy. These are traditionally the ethical issues seen to issue from a conception of data as information. Thick identity instead has to do with the experienced personhood of data subjects, and is a question of the way in which subjects are represented by data. These representations are a matter of ethical concern as they are the arena in which implicit personal and social values which inform moral judgments, are expressed and played out.

The paper then outlined what possibilities of handling data are made available by the digitisation, large-scale data-bases and other computational resources associated with

e-research. These include standardisation, reification, manipulation, and the analytical and interpretive capabilities of annotation, tagging, data-mining and visualisation.

In the face of the technological possibilities for handling digital data presented by e-research, researchers using data about human subjects have a further set of ethical issues to consider with regards to the representational content of data. These issues are complicated by the fact that it is not simply what – or who – is 'immediately' represented that needs to be considered, but also the implicit representational model or theory which goes some way to explaining why different people involved with data may feel that they have a particular stake in data. The third part of the paper outlined five different representational models: the naturalistic, isomorphic, figural, constructionist and isomorphic model. While not an exhaustive or definitive list, the models are put forward as a way of highlighting competing concerns and interests around data that can be traced back to the role of different roleplayers in determining representational content, as these may be perceived either by subjects or by researchers following different methodologies and research strategies.

The understanding of data as representation presents a challenge to the designers and users of e-research tools and technologies that go beyond the control of access and ensuring of confidentiality and anonymity. It is not possible to apply a cookie cutter model of ethical conduct with respect to Internet-enabled data relating to human research participants. Rather, it is important to try to understand what models of representation are at play for researchers and research participants, what negotiations are required around these models, and how technologies can be shaped to build in awareness of the ethical aspects of representations of people in Internet-enabled data.

References

Bamberg, M. (2004). Narrative discourse and identities. In J.C. Meister, T. Kindt, W. Schernus & M. Stein (Eds.), *Narratology beyond literary criticism* (pp. 213-238). Berlin and New York: Walter de Gruyter.

Barbaro, M. & Zeller, T. (2006). A face is exposed for AOL user 4417749. *New York Times*. 9th August, 2006. Retrieved January 15, 2008 from http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=1&adxnnl=1&oref=slogin&adxnnlx=1155744807DKt5sZmwY2uRl/mAsmGK9w.

Barthes, R. (1981). Camera lucida: Reflections on photography. New York: Hill & Wang.

Brey, P. (2000). Disclosive computer ethics. *Computers and Society, 30*(4), 10-16.

Burbules, N.C. (1998). Rhetorics of the Web: Hyperreading and critical literacy. In I. Snyder (ed.), *Page to screen: Taking literacy into the electronic age* (pp. 102-122). London and New York: Routledge.

Burk, D. (2007) Intellectual property in the context of e-Science. *Journal of Computer Mediated Communication*, *12*(2), article 13. Retrieved December 20, 2006, from http://jcmc.indiana.edu/vol12/issue2/burk.html.

Carlson, S., & Anderson, B. (2007). What *are* data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, *12*(2), article 15. Retrieved January 15, 2008 from http://jcmc.indiana.edu/vol12/issue2/carlson.html.

Corti, Louise (2000). Progress and Problems of Preserving and Providing Access to Qualitative Data for Social Research—The International Picture of an Emerging Culture [58 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* [Online Journal], *1*(3). Retrieved Janurary 15, 2008 from http://www.qualitative-research.net/fqs-texte/3-00/3-00corti-e.htm.

D'Agostino, G., Hinds, C., Jirotka, M., Meyer, C., Piper, T. & Vaver, D. (2006). On the importance of intellectual property rights for eScience and integrated health record. *Integrated Healthcare Workshop*. Edinburgh, UK.

Ess, C. (2002). Introduction. *Ethics and Information Technology 4*(3), 177-188.

Ess, C. & AoIR ethics working committee. Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee. Approved by AoIR, November 27, 2002. Accessed January 15, 2008 from http://www.aoir.org/reports/ethics.pdf.

Geertz, C. (1976). Art as a cultural system. *MLN 91*(6), 1473-1499.

Goguen, J. (1997) Toward a social, ethical theory of information. In Bowker, G.C., Turner, W., Gasser, L. (Eds.), *Social science, technical systems and co-operative work: Beyond the great divide* (pp. 27-56). Mahweh, NJ: Lawrence Erlbaum.

Gombrich, E.H. (1983). *Art and illusion*. Oxford: Phaidon.

Goodman, N. (1976). *Languages of art*. Indianapolis: Hackett.

Gross, L., Katz, J., & Ruby, J. (2003). *Image ethics in the digital age*. Minneapolis & London: University of Minnesota Press.

Habermas, J. (1988). *Le discours philosophique de la modernité*. Paris: Gallimard.

Heath, C. & Luff, P. (2000). *Technology in action*. Cambridge, England: Cambridge University Press.

Hinds, C., et al. (2005). Ownership of intellectual property rights in medical data in collaborative computing environments: *First International Conference on eSocial Science*. Manchester, UK.

Hyman, J. (1997). Words and Pictures. In Preston, J. (Ed.), *Thought and Language* (p. 42). Royal Institute of Philosophy Supplement: CUP.

Jirotka, M., et al. (2005). Collaboration and trust in healthcare innovation: The eDiaMoND case study. *Computer Supported Cooperative Work, 14*, 369-398.

Kindon, S. (2003). Participatory video in geographic research: A feminist practice of looking? *Area, 35*(2), 142-153.

Kluge, E.W. (2001). *The ethics of electronic patient records*. New York: Peter Lang.

Nissenbaum, H. (1988). Protecting privacy in an information age: The problem of privacy in public. *Law and Philosophy, 17*, 559-596.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*, 119-157.

Peacocke, C.A.B. (1987). Depiction. *Philosophical Review, 96*, 383-410.

Srinivasan, R. (2006) Where information society and community voice intersect. *The Information Society*, *22*, 355-365.

Strathern, M. (1999). *Property, substance and effect: Anthropological Essays on Persons and Things*. London & New Brunswick: Athlone Press.

Strathern, M. (2004). The whole person and its artifacts. *Annual Review of Anthropology, 33*, 1-19.

Tavani, H. (2004). Internet privacy: Some distinctions between Internet-specific and Internet-enhanced privacy concerns. *Ethicomp Journal 1*(2). Accessed January 15, 2008 from http://www.ccsr.cse.dmu.ac.uk/journal/abstract/tavani_h_internet.html.

Tomaselli, K. G. (1996). *Appropriating images: The semiotics of visual representation*. Højbjerg: Intervention Press.

Van Der Ploeg, I. (2003). Review of Eike-Henner Kluge. *Ethics of Electronic Patient Records, 5*(1), 66-67.

White, M. (2002). Representations or people? *Ethics and Information Technology, 4*(3), 249-266.