2008

Taking "Data" (as a Topic): The Working Policies of Indifference, Purification and Differentiation

Fletcher T.H. Cole
School of Information Systems, Technology & Management, University of New South Wales

Taking "Data" (as a Topic): The Working Policies of Indifference, Purification and Differentiation

Fletcher T.H. Cole
School of Information Systems, Technology & Management,
University of New South Wales

Abstract

The recent surge of interest in e-science presents an opportune moment to re-examine the fundamental idea of "data". This paper explores this topic by reporting on the different ways in which the idea of data is handled across many disciplines. From the accounts various disciplines themselves provide, these ways can be portrayed as the pursuit of three broad policies. The first policy is one of Indifference, which assumes the coherence of the data-concept, so that there is no need to explicate it further. The second policy is Purification, which identifies the essential characteristics of data according to the conventions of a particular discipline, with other modes systematically suppressed. The third policy allows for the Differentiation that is evident in the manifestations of data in various disciplines that utilise information systems. Greater appreciation among information professionals of the alternative approaches to data hopefully will enhance policy formulation and systems design.

Keywords

Data Management, e-Science, Information Systems Theory, Professional Discourse, Social Informatics

INTRODUCTION: ON MAKING DATA A TOPIC

The inexorable growth in the volume and complexity of digital data in many academic disciplines has led to a resurgence of interest in matters of their housekeeping. The advent of large-scale data-management projects in e-science and e-research indicates this renewed concern. A "veritable deluge of scientific data" is seen to engulf us (Hey & Trefethen 2005: 820; Borgman 2007a: 6, 115). For Borgman, the deluge is a pressing and daily reality in many fields, not only in "Big Science" but in less resource-hungry disciplines also (Borgman *et al* 2007b). The limits of what might be possible are currently a matter of intense discussion. In Australia, the response so far is to have serious reservations about the feasibility of any co-ordinated national approach (Buchhorn & McNamara 2006), but some views in official circles about the establishment of an Australian National Data Service are more optimistic (Australia. National Collaborative Research Infrastructure Strategy 2008).

Many of these data management projects adopt the point of view of an assets manager, and take data as a valuable resource, and as "capital" (Schroeder 2003). The possibility of the re-use of data is canvassed, immediately raising questions about interoperability. However, it would seem, as a prior requirement, we should review and clarify as far as possible our ideas about the commodity in question, namely "data".

Data is a mundane idea, but fundamental to disciplinary knowledge, as well as central to an understanding of information systems. Whether computer-related or not, an information system needs data to digest. So it is hardly surprising that, as the use of computers spread, various ideas about data quickly emerged in practice, almost to the point of anarchy (Machlup 1983: 646-649). However, interest has waxed and waned, with not a great deal of attention being paid to the concept of data as such, compared with that paid to its troublesome overlapping cousins, "information" and "knowledge" (Capurro & Hjørland 2003).

This paper aims to clarify some aspects of data, by reporting on three broad ways in which the characteristics of data are considered across a range of fields utilising computer-based information systems. These three "working policies", as I have termed them, are evident in the accounts that disciplines themselves provide of how they go about dealing with data. Sometimes the treatment of data, as a topic in its own right, is espoused in explicit policy statements. More commonly it is embedded within familiar methods – using terminologies, methods and work practices so familiar to a competent practitioner that much is hidden, for there is no need for them to be otherwise. Both the explicit and the implicit are the working policies-in-practice that are of interest.

The working conception of data has emerged in a multiplicity of distinct vocabularies of data-related terms, associated with various communities of practice and disciplinary traditions. The deployment of these vocabularies and their everyday use within disciplinary boundaries is part-and-parcel of the discipline, and normally regarded as unproblematic. However, there are occasions within a discipline which do render these

notions problematic, both in its history and in its day-to-day business. One of these occasions is when attempts are made to represent its data in terms manipulable in some computer-based information system. Typically the representation is resolved, and characterised (from the point of view of Information Systems and related disciplines) as an "application" of computational techniques to a "domain" – an "application domain" (Worboys & Duckham 2004: 19-20, 137; Elmasri & Navathe 2007: 31, 34, 59). The nature of these resolutions varies from discipline to discipline, but can be the source of ongoing and extensive debate, such as that within Geography about Geographic Information Systems (GIS) (Longley *et al* 2005).

The first course of action is associated with a working policy of *Indifference*. This is not so much ignoring the topic (although this happens), but rather assuming that there is no need for more than a cursory reference to it, nor any need to explicate the notion further. The other two policies do wish to explicate, but do so in different directions. The second of our policies proceeds in the direction of *Purification*. Here the essential characteristics of data are identified according to the conventional modes of representation of a particular discipline, and other modes are systematically rendered invisible. By way of contrast, the third policy works to identify the *Differentiation* that is evident in the various manifestations of data, and in the various Purifications that take place in practice. It provides an occasion on which to recover, and perhaps rehabilitate, that which otherwise might remain overlooked or forgotten, yet which may turn out to be crucial to the success of a systems design.

The different data policies discussed here are seen to arise out of the everyday affairs that constitute the life of a discipline and practice community. My account is therefore more an anthropological than a philosophical reflection, describing more how data is actually thought about, rather than prescribing how it should be. The field-work method has been literary, to gather and compare readily available accounts, in diverse fields where the issues are heightened sufficiently to identify these policies. It is written having in mind the tradition set by recent social studies of technology and workplace studies (Heath & Luff 2000; Hutchins 1996; Mackenzie 1996, 2001; Suchman 2007). It is a preliminary orientation which aims to note what counts as data, and how it counts.

INDIFFERENCE TO DATA

The first policy is one of *Indifference* to taking data as a working-concept or topic. In accord with well-established conventions and expectations, data are simply assumed to be there. Fundamental re-consideration is hardly thought to be necessary for accomplishing intelligible work outcomes. The topic is not so much "absent", as in no need of acknowledgement or elaboration, as "seen but unnoticed" (Garfinkel 1967: 44). This working policy is manifest in a number of ways, from exclusion, in not making any data-related reference at all, to more complex and ambiguous manifestations.

An instructive parallel to the argument of this paper which, somewhat ironically, illustrates this policy of Indifference, is to be found in Orlikowski and Iacono (2001). The burden of the authors is to point out that in the majority of articles appearing in *Information Systems Research* during the 1990s, "IT artefacts are either absent, black-boxed, abstracted from social life or reduced to surrogate measures" (130). Any serious treatment of Information Technology (as such, as a topic) was judged to be absent in 25% of the articles surveyed. Much of the treatment of IT in the remainder "draws on commonplace and received notions of technology, resulting in conceptualisations of IT artefacts as relatively stable, discrete, independent, and fixed" (121).

However, in setting out the different views of "Technology", only a brief mention is made of the "Information" of "Information Technology", and only in the discussion of the 8-9% of articles that took IT as an information processing tool. In this variation of the "tool view" of IT, it is "argued that what technology does best is to alter and enhance the ways that humans and organizations process information" (124). Information is seen to "flow" through organizations, and to accumulate in searchable "repositories". The central purpose of computerisation and IT is to enhance this reticulation. This is the closest the article comes to referring to data – for the purposes of the authors' argument it is a matter of Indifference. However, the parallels between the working policies with regard to the "IT artifact" and to "data" are striking, as will be mentioned in the concluding comments.

However, Indifference succeeds only to some extent, for there are occasions which do provoke some attention being paid to data, in its own right. This may happen when there are changes in disciplinary pre-occupations or computational capabilities. For instance, when there is proposal to merge data sets originating in different computational environments (where quite subtle differences can be irritatingly unsettling); or when attempts are made to coordinate data use across massive information systems (e.g. as in grid computing).

On these occasions data is referred to, but typically characterised as a single entity – unified and coherent – possibly as an IT artefact which is "relatively stable, discrete, independent, and fixed" (Orlikowski and Iacono 2001: 121). The work of resolution which has taken place, and routinely takes place to create that coherence will, at best, be acknowledged in passing, but more generally will be invisible. The analysis remains at this level of singularity, by referring to data as a unified entity in general, and by not explicating the data-concept further. Attempts might be made to categorise data in various ways, but with the data-concept remaining firmly intact

within those various categories. Certainly, no inclination towards deconstruction is evident. One prime example of this is when data is taken be a commodity.

Data as commodity

To all intents and purposes, a policy of Indifference is adopted when data is characterised as an undifferentiated commodity – that is, taken as a form of economic "goods". Arzberger *et al* (2004a, b), throughout their reports in *Science* and the *Data Science Journal* about work done for the OECD on international data access, consistently take data this way, such as in the following:

> Open access to publicly funded data provides greater returns from the public investment in research, generates wealth through downstream commercialisation of outputs, and provides decision makers with facts needed to address complex, often transnational, problems. (Arzberger *et al* (2004b: 1777)

Putting aside consideration of the merits of these claims, the working policy with respect to data is one which assumes coherence – uniformity even. This is understandable, in order to make high-level policy recommendations possible. To explicate the data-concept further is impractical (e.g. insufficient space), and unnecessary. The reasonable assumption is that readers will know something of the commodity being discussed, in order to focus attention on the management of that commodity. The overall effect is to discuss data, yes; but in accord with a policy of Indifference to the possibility of explicating the data-concept further. Paradoxically, at the same time data is being pointed to, it is simultaneously being pointed away from.

A more ambiguous example can be seen in Christine Borgman's recent book on building e-research infrastructure, *Scholarship in the Digital Age* (2007a). In it the basic concern is to explore the possibilities of building an infrastructure for data similar to that which exists for publications, which is seen as a "remarkably stable scholarly communication system" (xviii). In it, data (as a concept) is generally treated as an undifferentiated entity, but with some recognition of its mixed and hybrid nature. Most of the discussion characterises data as commodity, classifiable in a number of ways (e.g. by format, by discipline, by degree of evaluation). This treatment gives rise to a number of social and infrastructure policy issues, such as intellectual property, collaboration, storage and access, and funding. At one point it is noted that a sociological viewpoint might see the value of data not so much as "goods" (a term for economists) but value in terms of its place within the "variety and flexibility of social networks" (166). Here the spectre of Differentiation (to anticipate its appearance in our discussion) is raising its head, and with good reason, of which Borgman is aware.

For in her exposition there is also a tantalising counter-theme which indicates that, when compared with the everyday realities of data practice, a uniform and homogenous data-concept, and a working policy of Indifference to it, may need some re-examination. For while data might be taken as discrete artifact, we also notice that "their use is embedded deeply in the day-to-day practices of research" (xviii), and that these practices indicate profound variation. The data which may be "background context to one researcher" may be to another the "focus of research" (41). Different partners in a collaboration may identify data in quite different ways:

> For example, to the engineer building a sensor network, a reliable stream of bits may be data. Until the sensor network can detect biological phenomena in a reliable way, those bits may not be scientific data to a biologist. (183)

At several points in the exposition there is acknowledgement that data are highly contextualised in and by local practice, some of which can be explicitly documented and some of which cannot, because it is "enshrouded in tacit knowledge" (173) (c.f. 128, 166, 188, 197) or by their very irreproducible nature, as specimens, samples, or live animals (183). As such they are "subject to interpretation; their status as facts or evidence is determined by the people who produce, manage, and use those data" (121) and reference will be needed to those people if re-use of the data is contemplated (198). As a consequence, it is local "informal community-based quality-control mechanisms", rather than conformity to high-level de-contextualised standards, that play the crucial role in data evaluation (135). The not unreasonable conclusion is that this malleability in the significance and interpretation of data represents one of the "greatest challenges in building data repositories" (173). However, overwhelmingly, Borgman resists moving away from a unified data-concept. This suits the discourse of high-level policy formulation which is the purpose of her argument, but it also illustrates the tension between two different viewpoints about, and approaches to, the study of data (as a topic).

A similar resistance to Differentiation is observable in Wouters and Reddy (2003), and in the other papers in the volume in which it appears. The contrast here is stark, between the predominant portrayal of data as an unalloyed commodity (as capital) (Schroeder 2003), and actual data-related practice, a few accounts of which are also included, such as of collaborations in high-energy particle physics (at CERN). In these collaborations data has several manifestations, and is the prime subject of lengthy negotiations over its composition and permissible

interpretations (25-27). As will also be reported later, it involves hard work for the physicists wishing to make coherent sense of the data at hand. However, this more complicated picture does not disturb the undifferentiated view of data in the general argument here.

What can disturb the picture is when a more than cursory look at data (as a topic) is required. For this a new course of action needs to be undertaken, with a different working policy, one in which more detailed and precise explications of data can be explored. In practice, two alternatives appear to be pursued, and they proceed in opposite directions.

DATA PURIFIED

The first alternative moves to place strict limits on the concept of data, and in any analysis of it, to confine forms and modes of representation to that familiar to a particular disciplinary group. In this process of *Purification* or distillation ("reduction" is perhaps too severe a term for this) much of the basis for those representations which lies in other outside sources is overlooked, or consciously bracketed off, "put to one side, for the time being". Underlying this move is the concern to preserve the unified nature and coherence of the data-concept.

One does not need to look too far for examples. In database software one might readily think of the way data is distilled into the basic data types of SQL-99 and its variants. Numeric data can be shaped as integer numbers of various sizes (`INTEGER` or `INT`, and `SMALLINT`), or as real numbers at various degrees of precision (`FLOAT, REAL, DOUBLE PRECISION`) or with specific formats (`DECIMAL(i,j)`, `NUMERIC(i,j)`). Time is made up of `HOUR, MINUTE, SECOND`, and decimal fractions of a second as `TIME(i)` (Elmasri & Navathe 2007: 246-247). Behind the SQL standards there is the forgotten history of their formation, and of their precise implementation in electronic circuitry (aka hardware) which is also the outcome of a sometimes controversial history (Mackenzie 1996).

A more domestic example can be found in the various specifications of "postal address" which can be found built into Web-page forms, in order for customers to supply their contact details. Surprisingly often, postal codes and post office box addresses seem to present an irritating stumbling block, resulting in strange formatting or outright rejection of them by the controls embedded in the form. In these instances the attempts at Purification are not only noticeable but are ridiculous.

Measuring the River Tiber

A more elaborate illustration of the process of distillation, from the parallel world of GIS theory & practice, is unpacked by Worboys and Duckham (2004: 154-155). It concerns the stages a GIS might go through in order to calculate, for instance, the length of the River Tiber which lies within a specified distance of the Colosseum, in Rome.

Starting with a NASA satellite image of the area, the observation is immediately made:

> This data is not well suited to answering our question: it does not explicitly represent objects like "the Colosseum" or the "River Tiber," and contains much more detail than we need.

The image data is transformed so that it does become "well suited to answering our question" by highlighting the "relevant parts of Rome" in terms of a "model" consisting of geospatial "objects": "a **river** object (with "Tiber" as its name) and a **building** object (named "Colosseum")". There is also another, less prominent, geospatial object to be mentioned, namely, "the region within a 2.5-km radius of the Colosseum".

This is a start, and the next distinct stage is to purify these geospatial objects further, by identifying the means by which they might be represented geometrically. A set of geometric objects is drawn on to achieve this. An "arc" will represent the river object, the nearer eastern bank when we re-examine the satellite image, depending on how precise we want to be; a "point" for the building; and a "circle" describing a "disk" for the region.

The calculation procedure is then to find the length of the arcs which intersect with the disc, and sum them. Well, nearly, for there is further stage to be mentioned, and a number of other stages, which relate to how the machine does its sums, which are not mentioned. The objects and operations described so far

> … are not suitable for computation, because they are continuous and infinite. A process of discretization must convert the objects to types that are computationally tractable. For instance, a circle may be represented as a discrete polygonal area, arcs by chains of line segments, and points embedded in some discrete space. Operations such as **intersection** and **length** may then be computed using standard algorithms from computational geometry (155).

Even for a seemingly simple problem, the degree of refinement of the data in order to get it into a suitable shape for computation is considerable. However, a chain of arcs can hardly do the Tiber justice. Like the Mississippi,

"He jus' keeps rollin' along", with varying and complex width, and depth, and impeding structures, such as weirs, and all manner of inhabitants, and ingredients of varying levels of toxicity, which at some time may all need to be accounted for, in data, but which on this occasion are systematically excluded under a policy of Purification.

Distilling Provenance

Quite ambitious programs of Purification can be attempted as, for instance, in the recent EU-sponsored research into documenting the "provenance" of electronic data. The aim here is to redress the general problem that "electronic data does not typically contain historical information that would help end users, reviewers, or regulators make the necessary verifications" of it (Moreau *et al* 2008: 54). Accompanying data with a description of the process that led to its production, its "provenance", is proposed as a way of compensating for this lack.

The proposal has provenance being captured in process documentation of sufficient detail as to allow end-users, ideally, "to reproduce their results by replaying previous computations, understand why two seemingly identical runs with the same inputs produce different results, and determine which data sets, algorithms, or services were involved in their derivation" (54). The means by which this is to be achieved is to adopt the language and software design philosophy of Service-Oriented Architecture (SOA) and to document three types of assertions about data provenance within software services, "p-assertions", and the relationships between them at any point in time.

Interaction p-assertions describe "the contents of a message by a service that has sent or received it" (55), and "denote data flows between services" (56). *Service-state* p-assertions document "its internal state in the context of a specific interaction" (56). The characterisation of this may be very mixed, for it "may include the amount of disk and CPU time used by a service in a computation, the floating-point precision of the results it produced, or application-specific state descriptions" (56). *Relationship* p-assertions denote data flows within services, by describing how a service "obtained output data sent in an interaction by applying some function or algorithm to input data from other interactions" (55). The various "causal and functional data dependencies" can be tied together in a "directed acyclic graph", the DAG thus constituted assuming the status of a "core element of provenance representation" (56).

It is acknowledged that the "full detail of everything that ultimately caused a data item to be what it is could be quite large" (56), for instance, when properties of materials, settings of instruments and software are included, so that it is inevitable that some selection of possible indicators of provenance needs to take place. Documenting the "full detail of everything", or of "anything", is an unlimited project in practice – there is always more one can report. Process documentation is an *account* of process, and as with all accounts, it is assessed in terms of adequacy for the practical matters at hand. Not everything can be said, or needs to be said, in so many words (Garfinkel 1967). The proposal here for managing the provenance of electronic data, is to work in accord with a policy of Purification. Provenance is couched in terms of structured sets of p-assertions, and this purifying framework provides the means by which the "everything" that might be documented about provenance in a particular instance is distilled into provenance data.

For instance, in a hospital application for the management of organ transplants a fragment of the provenance record might be represented by the data item (graph node) "Donation decision" having the relationship (edge) of being "based on" to each of (nodes) "Blood Test Results", "Donor Data", and "Patient Death Notification" (55, 57; also Deora *et al* 2007).

The approach appears somewhat akin to the program of "organisational semiotics" promoted by Ron Stamper and colleagues, in which relationships between "signifier" (the p-assertions?) and "signified" in information systems are structured and formalised in terms of a denotative mapping of "complex expressions" onto "complex affordances" or "invariant repertoires of behaviour". The explicit but somewhat ambiguous assumption is that "words in different languages mapping onto the same affordances have the same meaning in the sense that their users would interpret them in the same behavioural manner" (Stamper 2001: 161-162).

It will be interesting to learn how well this proposal will work out, in the light of the massive and intractable "complex affordances" that medical practitioners have to negotiate, and given the chequered history of attempts to rationalise medical practice in information systems (Berg 1997). Choices as to what to highlight, and how to highlight it in the filtering process, are crucial. There are often counter-intuitive but "good organizational reasons" for bad records (Garfinkel 1967), and *ad hoc* undocumentable practices which are foundational in understanding much recorded data, such as in the reading of medical images (Slack *et al* 2007). Equally, there can be profound misunderstandings about the nature of the data, which can lead to system designs that in effect rely on "bad" organizational reasons for producing "good" but unworkable clinical records (Berg 1996; Heath & Luff 2000).

In addition, even in such a concerted effort to make "provenance" explicit, much remains unspoken, and treated with the Indifference discussed in the previous section. Yet even that which is unreported, apparently "unnoticed", is nevertheless "seen" (Garfinkel 1967), and forms an integral element of the process of defining and constructing provenance data. For instance, in reports of the provenance project, there does not seem to be any mention of how the graphs produced in the case studies are implemented in the software. Typically graphs are represented as matrices or lists, in order to render them in software, as arrays, and there are different ways this can be implemented at the various software levels – but to confirm this is not easy (if ever we should so want to).

Of course, this is not simply the Purification of the provenance data alone, for it is also an element in the Purification of the data whose history it is documenting. The provenance of data is logically and organizationally linked to the data whose history it is, not matter where it is stored. As provenance of data is an aspect of that data, to document it in the way proposed is also part of the process of purifying or "disciplining" that data (Berg 1997: 141-144). The provenance framework also provides the basis for preserving the coherence of the data as whole.

Yet, while the proposed solution admits to following a policy of Purification, the admission that some record of provenance, hitherto excluded, now needs to be made, would seem to be in tune with a policy of Differentiation. This suggests, as it did in the previous section, that in any particular case the dominant policy which is manifest in practice, in the practical decisions that are made, can have variations, and be ambiguous and mixed. All three working policies can and will be resorted to, simultaneously, but to varying degrees.

The extent to which the process of Purification is achievable without losing intelligibility makes for an intriguing topic, with immediate practical implications for information systems design. The unresolved issue is accounting for how, and when, and through whom, that which is suppressed is restored or compensated for, as it must, in the fullness of real-life, mundane, day-to-day practical affairs, in order to get any particular system to "work". The process of elimination, itself, can sometimes be a means of systematically accounting for what potential data (of a phenomenal field) has been left out, as in the stages identified in the GIS example above. The modelling of the Tiber, first as a *geospatial* object, can be distinguished from the subsequent *geometric* representation of that object (as an arc), as can the arc be distinguished from the *software* array that probably represents the vector matrix for that arc. From a different policy viewpoint, these stages could be taken as the places at which Differentiation in "the same data" can be seen. It is to this alternative working policy we now turn.

DIFFERENTIATION IN DATA

The second move (towards the explication of data-as-a-topic) works in accord with a policy of identifying the lines of juncture in the formulation of data, instead of struggling to hide the cracks. The analysis embarks on a path of deconstruction, more particularly, adopting a mode of *Differentiation*, at least to start with.

Differentiation as a working policy is evident in some research which aims to closely examine descriptions and other accounts of information systems (e.g. of GIS). Typically in the accounts under examination, seemingly heterogeneous data terminologies from different disciplinary sources are simply juxtaposed, in effect to form a *hybrid* conceptualisation (Cole 2005). Coherence in (and intelligibility of) the data-concept is assumed and unexplicated by the producers of these accounts – a matter of Indifference to them. Just how coherence is achieved, when confronted by a differentiated, heterogeneous and de-centered concept of data, becomes a major topic when pursuing such a path of Differentiation. An indication of how data coheres in everyday practice is suggested in a number of studies, often historical and anthropological studies where comprehensiveness and coherence have to be constantly juggled in writing the account, and it is often indicated inadvertently.

One rewarding source is Peter Galison's book *Image and Logic* (1997), about the development of experimentation in 20th-century particle physics. One theme that emerges is the extensive deployment of electronic computers and computing, along with the modes of data analysis that accompanied that deployment. In Galison's account, data can also be seen as a *hybrid*, being at the confluence of various discipline and practice communities: viz. image & logic traditions in experimental particle physics, with their long histories of interaction with a range of engineering and managerial practices, and with the more recent exploitation of computing and the simulation techniques enabled by them.

Coherence in the experimental enterprise and, by implication, the coherence of the data produced within and by that enterprise, is achieved and maintained within an on-going process of *negotiation and exchange* among these various discipline and practice communities. Galison points out the various locations, or "trading zones" (borrowing from culture-contact anthropology), in which such exchanges take place; some zones as small as a blackboard (as when young Julian Schwinger, quantum physicist, developed equivalent circuits for microwave engineers in the dead of night at MIT during WWII), or some as widespread as The Internet.

Mark I

One notable example of a "trading zone" reported by Galison is in the design and management of the Mark I Solenoidal Magnetic Detector in the 1970s. Mark I brought together three teams, one from the Lawrence Berkeley Laboratory, University of California (LBL), and the other two from the Stanford Linear Accelerator Center (SLAC).

LBL and SLAC represented two distinct traditions of experimentation and data analysis. For LBL data was primarily a visual record, a picture image, typified by a streak in a bubble chamber indicating the passage of a high-energy particle. However, for SLAC data were numbers, logical counts of physical events registered by specially-designed detection devices, such as a Geiger-Müller counter or a spark chamber.

Mark I satisfied both traditions by producing both pictures and pulses. The same data produced by Mark I was used by both, but it was very far from "raw", being the result of carefully designed computer-controlled "pre-processing" to screen out the noise of "background events" (PASS-1 software) and to smooth and classify trajectories (PASS-2). Its "sameness" was of a very particular sort, and the result of extensive negotiation between the two schools of thought. Self-evident it was not. This processed data was then used as the starting point by the different research groups, for their very different styles of analysis – hand-scanning in the case of LBL. Overall, a long-standing dream was realized; "the SLAC team saw Mark I as a full-bore electronic device with sensitivity in nearly every direction; and the LBL group saw the device as the realization of an electronic bubble chamber" (Galison 1997: 532).

The convergence did not happen by accident. As different instruments, each with in-built computing, needed be used in highly co-ordinated ways, so the corresponding need arose to co-ordinate the computing itself, more particularly the software and the data collection and analysis that needed to take place. This led in Mark I to the re-organization of research teams, or at least re-organization in their patterns of communication, affiliation, lines of authority. Up to this point "… most of the collaborators has been affiliated with the production of only one piece of hardware and its attached software for extracting quantitative information. Now that isolation had to end" (Galison 1997: 526).

Segregations, between entire institutions, present in the building of hardware, were breached. Rather, collaboration was re-organized in terms of the logical software characteristics associated with each device. To oversee it all, a super-group was created to supervise a program of integration in which "software would take output from the component software and weave it into a coherent data set with a clear and consistent set of calibrations" (Galison 1997: 527). This gives the idea of "Purification" a new twist, but from the alternative point of view of Differentiation it can be taken as hybridisation occurring at multiple levels and in several locations (including in and through the structuring of the data). The basis for the data-hybrid lay in the devices themselves and the organizational and experimental traditions which gave rise to them. The structure of the data reflects that genesis.

It also reflects the organizational and individual effort involved in creating that coherence, and in continuing to maintain it. That effort involved numerous attempts to bring the two data traditions and several other communities of practice closer together. The effort can be readily characterised as a process of negotiation and exchange among a number of interested parties, each with diverse understandings of data, each representing different histories, different preoccupations, intellectual traditions and material resources.

Monte Carlo

Coherence formation and maintenance within a trading zone is also evident in the introduction of Monte Carlo simulation techniques into particle physics. These have come to be an essential ingredient in particle physics including, in relation to the data, at the point of simulating the behaviour of specific devices, and "mimicking the microphysical events to extract process, clean and interpret data" (Galison 1997: 689) in order to produce workable results.

Monte Carlo simulation techniques grew out of the work in nuclear weapons design, particularly post-WWII, and the H-bomb. Their use quickly became widespread, in particle physics and in a number of other fields, and developed further, particularly in the context of a series of conferences during the 1950s. These brought together participants from a diverse range of disciplinary traditions, "pure mathematician, applied mathematician, physicist, bomb builder, statistician, numerical analyst, industrial chemist, numerical meteorologist, and fluid dynamicist" (Galison 1997: 752). Each had, however, and continued to have, a different take on its significance, often at precise points. As Galison reports:

> Dig down in the coding of a Monte Carlo and you find random numbers, perhaps just a line of code, 'random x', calling up such a list. But there is, within that superficially simple line of

> FORTRAN, some consequential philosophy about the very meaning of Monte Carlo, for 'random' meant different things to different groups of workers. (Galison 1997: 709)

These "meanings", including those about what counted as data, emerged out of on-going exchange, out of a "shared practice-based notion of Monte Carlo" (Galison 1997: 779); that is, residing in the agreed-to methods and procedures, not in any one disciplinary "interpretation".

There is both a structural and dynamic side to this. Data is viewed as an object "entangled" in that exchange; but also data is rendered intelligible and coherent in the negotiation that forms the basis of that exchange. The data-concept on its own, standing alone, does not, cannot, hold. It seemingly needs to be an integral part of an ongoing conversation, an account of which, under a policy of Differentiation, needs to be provided. As such, the coherence which data is presumed to exhibit has been re-specified.

CONCLUSION AND IMPLICATIONS

The main outcome of attempting to locate conceptualisations of data in the accounts of those who handle it is to conclude that the concept-in-use is not one, but many. It reveals itself as open to multiple and continuing interpretation through its deployment in different contexts, at different levels of abstraction, and in line with various working policies. There are a number of distinct ways of characterising data, and we can describe this variety in terms of three major policies or strategies associated with different courses of action and different series of collective judgements.

In accord with one of Orlikowski & Iacono's normative "premises" about IT artfacts, these policies indicate that our ideas about data can be "neither fixed nor independent, but they emerge from ongoing social and economic practices" (2001: 131). This suggests that perhaps we should see IT artifacts as being subject to similar policies in the ways they too are conceptualised. However, in contrast to the other premises that Orlikowski & Iacono recommend, we need not be surprised or disappointed by the differences. Rather, they represent alternative practical resolutions of the problem of maintaining the coherence and integrity of complex and differentiated socio-material phenomena. They provide intelligible courses of action in grappling with material cultures which incorporate information systems.

The first policy of Indifference, in its treatment of data as an indissoluble artifact, as "natural, neutral, or given" (Orlikowski & Iacono 2001: 131), preserves coherence by assuming it. This policy is useful when there is no need to explicate the data-concept further. In these situations it really might make sense to render data "either absent, black-boxed, abstracted from social life or reduced to surrogate measures" (Orlikowski & Iacono 2001: 130). In other situations ignorance of the nature of the data, and how it is being dealt with by an information system, is not bliss, but a recipe for disaster. Neither does the cursory treatment of data typical of our current information systems textbooks help much with cultivating in our students a more than Indifferent appreciation of its complex nature.

The second policy, Purification, deals with coherence by imposing it. A limited set of conventions, acknowledged by a particular community-of-practice, is privileged – all other modes are effaced. This policy is effective when the conventions are also understood and acceded to by those not belonging to that community, but who nevertheless have a significant "stake" in how the Purification takes place (e.g. "the user"). In the Information Systems world, there are numerous examples where clearly this has not happened, being reported under the rubric of "failure" in a burgeoning literature. When Purification is effective, it is interesting to investigate precisely why. For it may be that the doctrine has not been too-rigorously pursued, and that an awareness of that from which the data has been purified is still playing a significant role in making the system work.

Something similar has been pointed out in addressing Tony Hoare's question of "how did software get so reliable without proof?" (MacKenzie 2001: 302-303). Suggested reasons have included the use of systematic analysis and design methods which mitigate the risks of *ad hoc* "unstructured practices" that lead to unintelligible and unpredictable software. Alternatively, anticipation and avoidance of errors and problems play a role, as when users of systems compensate for or "repair" software errors and system output; e.g. by cross-checking output, by re-booting, by consulting a workmate or, if you are a pilot, by simply looking out of your cockpit window.

Purification can be a means to understanding data as a differentiated concept. The focus on distillation, can require a substantial, if not equal, awareness of the impurities being removed. Keeping track of the impurities, and not discounting them, provides a resource when the purified data is put to use; when that which is suppressed is compensated for in the undifferentiated fullness of day-to-day practical affairs.

The third policy, Differentiation, confronts coherence, and denies it. The integrity of data is to be understood in the same light as Orlikowski & Iacono want us to see the IT artifact, as

> … a multiplicity of often fragile and fragmentary components, whose interconnections are often partial and provisional and which require bridging, integration, and articulation in order for them to work together (2001: 131).

But work together they do. The presumed fragility may be an exaggeration – for data is no more fragile than any other socio-material entity which may "emerge out of social and economic practices". For while the data-concept is de-centered, shifting, and "dynamic", in its various manifestations it is firmly "embedded in some time, place, discourse, and community". The challenge under a policy of Differentiation is to know when to stop identifying the "components" and tracing the "interconnections". That, of course, "will depend" – on the purpose, and the situation, and when sufficiently intelligible "bridging, integration, and articulation" has been done, so as to present data as an integrated concept. It will be coherence re-specified.

While three distinct policies are evident in the practice of data, they are not independent. Purification summons up Differentiation. Differentiation, through identifying aspects of data, facilitates Purification. Both provide an alternative means through which Indifference can be addressed. But in normal everyday affairs it will be back to Indifference that they both must return, in which data is seen, taken for "what it is", and without need of much comment.

The policies weigh differently in different situations. They each shape expectations as to how the notion of data can be made to work. In pursuing a policy of Purification, careful judgement is needed in assessing what is baby and what is bath-water. Within a policy of Differentiation, careful judgement is also needed as to salience; not everything can or needs to be said "in so many words". At least an awareness of these three policies at work means we can become a little more sophisticated in taking, when we take, "data" for granted.

REFERENCES

Arzberger, P. *et al* 2004a. "Promoting Access to Public Research Data for Scientific, Economic, and Social Development," *Data Science Journal* (3), 29 November, pp 135-152.

Arzberger, P. *et al* 2004b. "An International Framework to Promote Access to Data," *Science* (303: 5665), pp 1777-1778.

Australia. National Collaborative Research Infrastructure Strategy. 2008. *Platforms for Collaboration*. http://www.pfc.org.au/bin/view/Main/AeRIC (accessed 8 Sept 2008)

Berg, M. 1996. "Practices of Reading and Writing: The Constitutive Role of the Patient Record in Medical Work", *Sociology of Health and Illness*, (18:4), pp 499-524.

Berg, M. 1997. *Rationalizing Medical Work: Decision-Support Techniques and Medical Practices*. Cambridge, MA: MIT Press.

Borgman, C.L. 2007a. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet.* Cambridge, MA: MIT Press.

Borgman, C.L., Wallis, J.C. and Enyedy, N. 2007b. "Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries," *International Journal on Digital Libraries* (7:1), pp 17-30.

Buchhorn, M. and McNamara, P. 2006. *Sustainability Issues for Australian Research Data: The Report of the Australian e-Research Sustainability Survey Project (AERES Report).* Canberra: Australian Partnership for Sustainable Repositories (APSR), The Australian National University. Retrieved 26 June, 2008 from http://www.apsr.edu.au/aeres/

Capurro, R. and Hjørland, B. 2003. "The Concept of Information," *ARIST: Annual Review of Information Science and Technology*, (37), pp 343–411.

Cole, F. 2005. "The Discourse of Data: Exploring Data-Related Vocabularies in Geographic Information Systems Description," *Journal of Information Science*, (31:1), pp 44-56.

Deora, V. *et al* 2006. "Navigating Provenance Information for Distributed Healthcare Management," IEEE/WIC/ACM Web Intelligence Conference. Retrieved 21 June, 2008 from: http://www.gridprovenance.org/publications/vdeora-provenance.pdf.

Elmasri, R. and Navathe, S.B. 2007. *Fundamentals of Database Systems.* 5th ed. Boston: Pearson, Addison Wesley.

Galison, P. 1997. *Image and Logic: A Material Culture of Microphysics.* Chicago: University of Chicago Press.

Garfinkel, H. 1967. *Studies in Ethnomethodology*. Englewood Cliffs: Prentice-Hall.

Heath, C, and Luff, P. 2000. *Technology in Action.* Cambridge: Cambridge University Press.

Hey, T. and Trefethen, A. 2005. "Cyberinfrastructure and e-Science," *Science* (308: 5723), pp 818-821.

Hutchins, E. 1995. *Cognition in the Wild.* Cambridge, MA: MIT Press.

Longley, P., Goodchild, M.F., Maguire, D. and Rhind, D. eds. 2005. *Geographical Information Systems: Principles, Techniques, Management and Applications.* Abridged [augmented] 2nd ed. New York: Wiley.

Machlup, F. 1983. "Semantic Quirks in the Study of Information", In: Machlup, F., Mansfield, U., eds. *The Study of Information: Interdisciplinary Messages.* New York: Wiley, pp 641-671.

MacKenzie, D. 1996. "Negotiating Arithmetic, Constructing Proof," In: *Knowing Machines: Essays On Technical Change*. Cambridge MA: MIT Press, pp 165-183.

MacKenzie, D. 2001. *Mechanizing Proof: Computing, Risk, and Trust.* Cambridge: MIT Press.

Moreau, L. *et al* 2008. "The Provenance of Electronic Data," *Communications of the ACM,* (51:4), April, pp 852-858.

Orlikowski, W.J. and Iacono, C.S. 2001. "Research Commentary: Desperately Seeking the 'IT' in IT research – A Call to Theorizing the IT Artifact," *Information Systems Research* (12:2), June, pp 121-134.

Schroeder, P. 2003. "Digital Research Data as the Floating Capital of the Global Science System," In: Wouters, P. and Schroder, P., eds. *Promise and Practice in Data Sharing*. (The Public Domain of Digital Research Data) Amsterdam: Networked Research and Digital Information (Nerdi) NIWI-KNAW, pp 7-12. Retrieved 17 June, 2006, from http://dataaccess.ucsd.edu/PromiseandPracticeDEF.pdf

Slack, R., Hartwood, M., Proctor, R., and Rouncefield, M. 2006. "Cultures of Reading: On Professional Vision and the Lived Work of Mammography," In: Hester, S. and Francis, D., eds. *Orders of Ordinary Action: Respecifying Sociological Knowledge.* Aldershot: Ashgate, pp 175-193.

Stamper, R.K. 2001. "Organisational Semiotics without the Computer?" In: Liu, K. *et al* eds. *Information, Organisation, and Technology: Studies in Organisational Semiotics.* Dordrecht: Kluwer Academic Publishers, pp 115-171.

Suchman, L. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge: Cambridge University Press. (1st ed 1987)

Worboys, M.F. and Duckham, M. 2004. *GIS: A Computing Perspective.* 2nd ed. Boca Raton: CRC Press.

Wouters, P. and Reddy, C. 2003. "Big Science Data Policies," In: Wouters, P. and Schroder, P., eds. *Promise and Practice in Data Sharing*. (The Public Domain of Digital Research Data) Amsterdam: Networked Research and Digital Information (Nerdi) NIWI-KNAW, pp 13-40. Retrieved 17 June, 2006, from http://dataaccess.ucsd.edu/PromiseandPracticeDEF.pdf

COPYRIGHT