Human Data Interaction: Historical Lessons from Social Studies and CSCW

Andy Crabtree
School of Computer Science, University of Nottingham, UK.
andy.crabtree@nottingham.ac.uk

Richard Mortier
Computer Laboratory, University of Cambridge, UK.
richard.mortier@cl.cam.ac.uk

Abstract. Human Data Interaction (HDI) is an emerging field of research that seeks to support end-users in the day-to-day management of their personal digital data. This is a programmatic paper that seeks to elaborate foundational challenges that face HDI from an interactional perspective. It is rooted in and reflects foundational lessons from social studies of science that have had a formative impact on CSCW, and core challenges involved in supporting interaction/collaboration from within the field of CSCW itself. These are drawn upon to elaborate the inherently social and relational character of data and the challenges this poses for the ongoing development of HDI, particularly with respect to the 'articulation' of personal data. Our aim in doing this is not to present solutions to the challenges of HDI but to articulate core problems that confront this fledgling field as it moves from nascent concept to find a place in the interactional milieu of everyday life and particular research challenges that accompany it.

Introduction

> " … what experience and history teach is this - that people … never have learned anything from history, or acted on principles deduced from it. Each period is involved in such peculiar circumstances, exhibits a condition of things so strictly idiosyncratic, that its conduct must be regulated by considerations connected with itself, and itself alone." (G. W. F. Hegel)

We write this paper in the sincere hope that Hegel got it wrong, and that something of value may be learnt from historical works to inform considerations in the current circumstances of inquiry. The current circumstances and considerations we refer to are, as the title of this paper suggests, to do with Human Data Interaction or HDI, an emerging field of computer science concerned with understanding and developing the underlying technologies required to support human interaction with digital data (algorithms, analytics, visualisations, etc.). This field of research is driven by the recognition that digital data has become what phenomenologists (Hegel included) might call an 'object-in-itself', a distinctive phenomenon worthy of treatment in its own right as reflected, for example, in widespread current interest in 'Big Data'. Whether the data is big or small is not of particular concern here, however. Rather, the issue is essentially one of developing some appreciation of what the 'I' means in HDI, and the challenges this raises for the development of computer support for human data interaction in the round.

Our aim in this paper is, first, to understand what HDI is about as a distinctive field of research and to unpack the ways in which human data interaction is construed of. We then reflect on social studies of human data interaction to highlight the contrast between technological and ethnographic conceptions, which suggest that data is not so much a thing-in-itself as a thing-embedded-in-human-relationships (Star and Griesemer 1989). The contrast draws attention to the interactional demands that may be placed on HDI as a socio-technical infrastructure (Star 1999), and the subsequent interactional challenges that accompany this, particularly the development of interaction mechanisms that support the 'articulation' of personal data between parties involved in sharing it (Schmidt and Bannon 1992). Our reflections draw attention to the need to develop social models and mechanisms of data sharing that enable users to play an active role in the process. We identify a number of core research challenges involved in bringing this about, which revolve around personal data discovery, data ownership and control, data legibility, and data tracking.

What is Human Data Interaction?

As per most academic ventures the reader might anticipate that answering the above question won't be straightforward, and indeed even a cursory glance at the literature makes it visible that different meanings already attach to the term; we discern at least 5 distinct 'versions' to date.

- HDI is about human manipulation, analysis, and sense-making of large, unstructured, and complex datasets (Elmqvist 2011).
- HDI is about delivering personalized, context-aware, and understandable data from big datasets (Cafor 2012).

- HDI is about providing access and understandings of data that is about individuals and how it affects them (Mashhadi et al., unpub. manu.).
- HDI is about federating disparate personal data sources and enabling user control over the use of 'my data' (McAuley et al. 2010).
- HDI is about processes of collaboration with data and the development of communication tools that enable interaction (Kee et al. 2012).

While distinct, there is a sense of continuum running through the different versions of HDI; a connecting thread as it were that suggests a) that there is a great deal of digital data about, so much so that it might be seen as 'the next frontier' for computing and society alike (Pentland 2012); b) that HDI is very much configured around large amounts of 'personal data' (whether in terms of delivering personalised experiences or in terms of it being about individuals); and c) that interaction covers a range of interrelated topics from data analytics to data tailoring, and enabling access, control, and collaboration.

On reading the literature, such as it is (this is a fledgling field), we come to the view that HDI is not about data per se then, not even digital data, but is very much centred on digital data pertaining to people and digital data that may be considered to be 'personal' in nature. As McAuley et al. (2010) and Haddadi et al. (2013) put it respectively,

> "Modern life involves each of us in the creation and management of data. Data about us is either created and managed by us (e.g., our address books, email accounts), or by others (e.g., our health records, bank transactions, loyalty card activity). Some may even be created by and about us, but be managed by others (e.g., government tax records)."

> "An ecosystem, often collaborative but sometimes combative, is forming around companies and individuals engaging in use of personal data."

We are not suggesting that everyone who has used the term HDI buys into this conception, only that it captures a distinct problematic: the management and use of personal data in society at large (Mortier et al. 2014, Haddadi et al. 2015). HDI is a distinctively socio-technical problematic, driven as much by a range of social concerns with the emerging personal data 'ecosystem' as it is by technological concerns, to develop digital technologies that support future practices of personal data interaction within it. It will come as no surprise then that HDI has opened its doors to interdisciplinary engagement. In this paper we seek to go beyond considerations of social context – i.e., of the ethical, political, economic, and legal considerations that frame HDI (as important as these are) – to understand how the social might actually *shape* interaction within HDI. To do that, we first need to unpack how interaction is construed of within the field.

How is Interaction Construed within HDI?

In begging the question of how interaction is construed within HDI we recognise the need to move beyond the vague generalities outlined above and say something concrete about how analytics, tailoring, access, control and collaboration (etc.) are to be supported so as to afford human data interaction within the personal data ecosystem. It is perhaps useful to first consider the posited nature of interaction in HDI.

> "We are not dealing with explicit interactions but with more passive scenarios. In HDI we consider people interacting with apparently mundane infrastructure, which they generally do not understand and would rather ignore." (Haddadi et al. 2013)

HDI seeks to transform the current circumstances of interaction from a 'passive' situation in which personal data or more specifically 'data about you' is generated in your mundane interactions with digital infrastructure and is increasingly accessible to third party use, into an active situation in which 'my data' and its subsequent use is actively managed and controlled by the people who produce it. The interactional situation is further complicated by the recognition, as stated above, that 'data about you' may not be yours but may either be generated by you on behalf of a third party (e.g., the taxman) or by third parties (e.g., retailers) in your interactions with their services. This creates an interactional situation that negates 'data containment' – i.e., the idea that 'data about you' could be handed over to you and that third parties could be prohibited from distributing copies of it without your permission. That 'my data' can, in the current circumstances, be readily copied and distributed by third parties further compounds the problem of human data interaction.

So we have a current interactional situation in which 'data about you' is passively generated by you in your mundane interactions with digital infrastructure, and rather more actively by you on behalf of others or by others themselves in your interactions with their services, and that any of this data is, in principle and increasingly in practice, available to infinite replication and (re)distribution. The current situation underpins the very viability of the Big Data society, and the situation is on the verge of an exponential explosion as the Internet of Things locates a myriad data gathering objects in the fabric of everyday life. How, then, is the interactional situation now and in the foreseeable future to be addressed?

Mashhadi et al. (unpub. manu.) propose various models of human data interaction (ranging from the pay-per-use model to the data market and open data models), but provide little insight into how interaction would actually be provided for within them. By contrast McAuley et al. (2010) seek to provide computational means of supporting human data interaction through the development of 'dataware', which seeks to federate disparate sources of data 'about me' and to

build digital infrastructure that enables people to exercise control over the use of data that belongs to them and/or is about them. Data federation seeks to enable people to become involved in third party processing of personal data without requiring that they take sole responsibility for the data. Data control complements data federation by focusing on who is gathering, processing and distributing 'my data', when and for what purposes, and the means by which an individual can enable processing services and applications to access the data on their behalf.

The Dataware Model

The dataware model does not capture all there is about HDI, but rather provides a particular instantiation of some core concepts. The model is based on three fundamental types of interacting entity: *the user*, by or about whom data is created; the *data sources*, which generate and collate data; and the *data processors*, which wish to make use of the user's data in some way. To assist the user in managing the relationship between these entities, the model posits that the underlying technology will provide the user with a 'personal container' or 'databox', which will enable them to oversee and manage access to their data sources and processing of their data by various 'data consumers'. This is a logical entity formed as a distributed computing system, with the software envisaged to support it consisting of a set of APIs providing access to data held by data sources. Data processors would write code to use these APIs, and then distribute that code to the data sources which take responsibility for executing it and then return results as directed by the data processor. The final and key piece of infrastructure envisaged is a *catalogue*, within which a user would register all their data sources, and to which processors would submit requests for metadata about the sources available, as well as requests to process data in specified ways.

From a user's point of view, interaction with this model works as follows: processors desiring access to one or more datasets within the catalogue present a request for access along with information about the request (minimally a representation of the processing to be carried out); the user permits (or denies) the request, which is indicated by the catalogue returning some form of token to the processor representing granted permission; the processor subsequently presents the request (the processing to be carried out) and the token to the data sources it covers; finally, the data sources return the results of the processing as directed in the request to the data consumer. The model assumes that the catalogue and the data sources it references are governed by the user, including logging and auditing the uses made of data so that the user can retrospectively inspect what has been done, when, by whom and to what end.[1]

---

[1] The model also permits a user to operate multiple catalogues, independent of each other, thereby providing a means to control the problems of linking accounts across different sources. Interactions between such catalogues are not considered an explicit feature of the system.

Dataware can be seen as an attempt to build a digital infrastructure that supports human data interaction by surfacing a user's personal data sources and what third parties would do with them or have done with them. It construes of the 'I' in HDI as an *accountable transaction* between the parties to it, configured in terms of request, permission, and audit. In this respect it potentially transforms the current situation of interaction, which is characterised by the largely unaccountable use of personal data by third parties, but it leaves untouched how any such transaction will, in practice, be accountably conducted. We do not mean by this how requests, permissions and audits will actually be carried out. Though these are important matters to be resolved, what have in mind here are the accountable matters that any such actions turn upon and would have to turn upon if they were actually to be brought about and be 'pulled off' in the real world. It is towards unpacking what we mean by 'accountable matters' of human data interaction that we now turn.

Accountable Matters of Human Data Interaction

Human data, as any other data, might usefully be understood as a 'boundary object' (Star and Griesemer 1989), a common notion in CSCW where it has been used both to shape studies of cooperative activity and concepts of CSCW systems, particularly the notion of 'common information spaces' in which boundary objects are understood as 'containers and carriers' of information between actors and organisations (Bannon and Bødker 1997). It is not our intention here to provide a detailed review of Star and Griesemer's work, as this is a well-trodden path in CSCW, but rather to draw out some salient features of relevance to HDI. Although the concept of a boundary object originated in ethnographic studies of collaborative activity in science, it is possessed of features that are of broader relevance to understanding the nature of information or data in human interaction.

Star and Griesemer's original account of boundary objects can be read more generally to suggest that human data interaction turns upon 'a mutual modus operandi' involving 'communications' and 'translations' that order the 'flow' of information through 'networks' of participants. This, in turn, creates an 'ecology' of collaboration in which data interaction becomes stable. As stable entities boundary objects inhabit 'several intersecting worlds' (e.g., the individual's, the supermarket's, the bank's) and meet the information requirements of each. Your credit card receipt might be seen to be a boundary object – as well as detailing how much you spent on shopping at the supermarket, for example, it enables the supermarket to bill your bank, and your bank to clear payment for your goods. As Star and Griesemer put it,

> "Boundary objects … are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites … They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognisable, a means of translation."

Thus, the credit card receipt is a proof of payment for you, proof that a request for payment from your bank to the supermarket is valid, and proof that a valid transaction has been made on your behalf by the bank to the supermarket. The receipt spans several intersecting social worlds, has different meanings in each, and yet maintains a common identity across sites: it is a record of a financial transaction that coheres across social worlds.

The 'coherence' of boundary objects is something that needs a little unpacking. That boundary objects can cohere across social worlds, that they are *recognisable* and thus accountable to multiple parties, turns upon 'invisible work' (Star 2010) or action and interaction that largely goes unrecognised, is taken for granted and ignored. This invisible work is, nonetheless, consequential for the design of computational systems, as we have seen in other areas of systems development (Suchman 1995). The upshot is that when we turn to boundary objects we are not just turning to a 'container' or 'carrier' of informational material then, but to the interactional grounds upon which the 'containing' and 'carrying' of such material gets done. This means that boundary objects are 'the stuff of action' and are, as such, embedded in some underlying 'arrangement' of collaborative work (Star 2010). Boundary objects are inherently social then, cohering *in* action and interaction that inevitably reaches 'beyond a single site'. In this respect boundary objects are also spatially and temporally distributed, which points to their 'processual' character as well.

> "The object (remember, to read this as a set of work arrangements that are at once material and processual) resides between social worlds … where it is ill structured." (ibid.)

The 'ill-structured' nature of boundary objects points to their inherent malleability, though what is interesting here is how, over the course of being translated across social worlds and in 'tacking back-and-forth' between the local needs of parties to their collaborative production and use, boundary objects become 'well-structured' and stable. As Star (ibid.) puts it,

> " … when the movement between the two forms either scales up or becomes standardised, then boundary objects begin to move and change into infrastructure, into standards (particularly methodological standards), and into things and yet other processes, which have not yet been fully studied as such."

The coherence of boundary objects ultimately turns upon their standardisation, which is provided for 'methodologically' – i.e., through the development of methods for communicating data and coordinating data sharing. As these methods

become standardised they become part and parcel of the mundane 'infrastructures' that permeate everyday life.

Here we touch upon another major concept to emerge from Star's ethnographic work, and something that has 'not yet been fully studied as such' – infrastructure, or the study of 'boring things' (Star 1999). What she means is unremarkable things, taken for granted things, things that are invisible-in-use. Infrastructure is a familiar feature of everyday life:

> "People commonly envision infrastructure as a system of substrates – railroad lines, pipes and plumbing, electrical power plants, and wires … This image holds up well enough for many purposes – turn on the faucet for a drink of water and you use a vast infrastructure of plumbing and water regulation without usually thinking much about it." (ibid.)

Star suggests that there is more to infrastructure than the configuration of technology, of pipes and wires and power plants, etc. She suggests that it is also, and essentially, 'relational', and that it is by virtue of this that infrastructure comes to be embedded in the 'organised practices' of everyday life.

> "So, within a given cultural context, the cook considers the water system as working infrastructure integral to making dinner … Analytically, infrastructure appears only as a relational property, not as a thing stripped of use." (ibid)

The notion of boundary objects makes it clear that there is a great deal more to data, and the development of infrastructures to support interaction with it, than meets the eye. It makes it clear that data is, as Star puts it, an 'n-dimensional' social object, containing (1) informational material that is (2) distributed spatially and temporally across (3) participating sites through (4) processual arrangements of collaborative work that are (5) coordinated through standardised methods of communication, elaborating (6) particular contextual relationships that embed data (7) in organised practices of everyday life and (8) thereby constitute infrastructure. So what?

Boundary Objects, HDI and CSCW

The infrastructural view on boundary objects suggests that data is not so much an object-in-itself as it is an object-embedded-in-human-relationships, and that data transactions within those relationships are possessed of particular accountable properties or 'dimensions' that provide for the coherence of human data interaction. We might ask the question then, is the interactional arrangement request-permission-audit *sufficient* to make HDI into a mundane infrastructure? When considered from a socio-technical viewpoint on infrastructure, the dataware model, while marking a necessary step-change, would seem to lack overall coherence.

Take, for starters, the basic principle of human data interaction as elaborated by the notion of boundary objects: that it turns upon a 'a mutual modus operandi', which involves 'communications' and 'translations' that order the 'flow' of information through 'networks' of participants. At first glance it might appear that HDI within the dataware ecology reflects this principle, but in what sense is interaction mutual? Communications are driven by third parties, not by the people whose data is being transacted and translated. That you or I are implicated in the interaction through requests and permissions does not make it a mutual modus operandi. The 'user' (though this seems a strange term in this context, the 'used' seems more apposite) is essentially on the receiving end of interaction; it is something done to them, not by them. Even if they do have the ability to refuse or remove permissions, the user is dealing with one-way traffic. The dataware modus operandi is asymmetrical and begs the question of what a symmetrical relationship might look like, e.g., how might users drive data sharing by (for example) actively seeking out data processors?

Complicating the situation is the inherently cognitive character of the dataware model. It is a model based on 'my data' and on data 'about me'. It is, as such, an individuated model that ignores the n-dimensional character of human data. What we mean by this is that much of the data that 'I' generate is produced in 'my' interactions *with others*. Data is relational and it often relates not so much to 'me' or 'you' but to 'us', and with this the coherence of the 'my data' model starts to break down and break down in challenging ways. It is not just a matter of handling what, for example, 'you' posted on 'my' Facebook page, but of handling the media we produce and consume together. Thus, the unit of data is not always 'mine' but frequently 'ours'. How is 'our data' to be handled? How is social data to be catalogued and governed?

The social character of human data in turn raises serious issues of data ownership and control. The individuated model makes 'me' the owner and controller of data, but as this model breaks down in the face of the social, how is ownership and control of 'our' data to be provided for? It's not 'simply' a matter of enabling ownership and control over data that cannot be disambiguated and assigned to individuals, or enabling a self-defined cohort to pool or aggregate its members' data, such that, for example, one person in the home could 'house keep' personal data for all householders, much as we see with respect to the day to day management of the home network (Tolmie et al. 2007). A host of relational issues are wrapped up in any such endeavour: the age of members of 'our' cohort will shape ownership and control, as will the personal situations that members find themselves in. Who, for example, will own and control 'our' children's personal data? And what about elderly, infirm or temporally incapacitated members of 'our' cohort? Situated within a lively social context, and accompanied by differing relational rights and obligations, ownership and control cannot be permanently fixed and tied to an individual, as the dataware model

presumes, but will instead change over time with respect to a host of evolving relationships and contingencies.

The inherent rub between 'my data' and 'our data' will need to be managed too. Even were users able to manage a pool of 'our' data, there persists a tension between members with regard to what should be pooled and what should remain 'mine'. This raises problems both of ownership and control. Take, for example, a young child's personal data – who owns it and who controls it? It cannot be assumed that the same person exercises ownership *and* control. Ownership may well reside with the person to whom the data applies as it were, but control in such a situation may well be delegated to another (e.g., a parent) thereby reflecting current organised practices of personal data handling (take, for example, a young child's health records or bank details). The same does not apply to a teenager, however. As they develop their independence we might well expect, again in line with current organised practices of human data interaction, that they will assume control over their own data along with a great many other aspects of their life, though this may be a phased rather than a sharp transition. The same may apply in reverse to an elderly member of the cohort who wishes to hand over the running of their affairs to someone else.

The subtleties of human data interaction in the social world make ownership and control into complex matters in which 'my data' must co-exist alongside 'our data', and mechanisms must exist to enable translations between the two. There is, then, a need to develop a much more encompassing and dynamic model of human data interaction, including the possibility for users not only to refuse or remove permissions but also, to redact data, both internally within a cohort (whether it be a family or some other grouping of people) and externally in our interactions with third parties. In the real world data sharing is 'recipient designed' – i.e., shaped by people with respect to the relationship they have with the parties implicated in the act of sharing. What you tell people of how much you smoke or drink or what kinds of foodstuff you eat and how much you weigh, for example, very much depends upon who you are doing the telling to. It is well known by doctors, for example, that such matters are grossly underestimated when they are told to them. The same applies more generally; not that we grossly underestimate things but that we are selective in what we divulge about our personal lives, with the 'selectivity' being done with respect to our relationship to the other parties involved.[2]

These problems, which are be no means exhaustive of the challenges confronting efforts to build digital infrastructures supporting human data interaction, suggest that there is a strong sense in which we need to factor

---

[2] HDI construes of the recipient as *the processor*, which presents a particular request for computation to be carried out to the data source after it has been granted permission. While this hold true, the issue is to enable *the user* to design permission with respect to just what of the data is available to the processor, and to others within a particular cohort too. Recipient design draws our attention for the need to support human judgement, decision-making and intervention in the course of human data interaction.

'articulation work' into HDI. Like the notion of boundary objects, articulation work is a familiar concept in CSCW (Schmidt and Bannon 1992), where it typically refers to an important feature in the design of cooperative information systems for the workplace. While some may be inclined to argue that the workplace is all that it applies to, we think it may also usefully extend to human data interaction in the round insofar as there is a *necessary interdependence* between users, both as individuals in their own right and as potential members of self-defined cohorts, and third parties who would purpose their data. Wittingly or not the dataware model makes users part of a division of labour whose work involves the organised harvesting of personal data, whatever its purpose (whether to drive the delivery of personalised digital services to users, or for financial reasons by users, or by all parties involved for the social good, etc.).

This may be a contentious claim to make and it is worth briefly reviewing what is distinct about cooperative work to substantiate it, as it may be tempting to see the dataware user as someone engaged in individual activity rather than cooperative work. As Schmidt and Bannon argued many years ago, cooperative work is a distinct category of work having certain fundamental features irrespective of technology past, present or future.

> " … the conception of cooperative work … does not assume or entail specific forms of interaction such as mode and frequency of communication, comradely feelings, equality of status, formation of a distinct group identity, etc. or even specific organisational settings." (ibid.)

Indeed, Schmidt and Bannon go on to argue that cooperative work is not 'necessarily congruent' with the boundaries of formal organisations or legal definitions of work relations.

> "Cooperative work is constituted by interdependence in work, that is, by work activities that are related as to content in the sense that they pertain to the production of a specific product or service." (ibid.)

The necessary interdependence of actors defines cooperative work, without presupposition as to the formal or legal status of the relationship between the parties to it. Schmidt and Bannon thus suggest that the term cooperative work should be taken as a 'general and neutral designation' of multiple persons working together to produce a product or service. People may then be said to be engaged in cooperative work if they are mutually dependent upon one another in the production of a product or service. While essentially individuated, the dataware model nevertheless configures a relationship of mutual dependence between users and third parties who would purpose their data. The dataware user may not be employed by third parties in a formal or legal sense, and thus be deemed to be part of an organisation, but they are inevitably enmeshed in cooperative work.

A core feature of cooperative work is 'articulation work' – i.e., the meshing together of distributed individual activities (Strauss 1985). Drawing off Strauss, Schmidt and Bannon tell us that articulation work is a 'supra-type of work', an unavoidable 'overhead' implicated in the doing of any activity that is bound up with others. Someone taking a walk has to mesh the business of walking with those around them, for example, has to coordinate their individual actions with the other people whose paths they cross. Articulation work speaks to the coordinate character of human action, to the *gearing in* of individual courses of action with one another. It is done in innumerable and manifold ways, though Schmidt (1994), drawing off a range of ethnographic studies, highlights several generic features of action and interaction that coordination turns upon. These include 'maintaining reciprocal awareness' of salient activities within a cooperative ensemble; 'directing attention' towards the current state of cooperative activities; 'assigning tasks' to members of the ensemble; and 'handing over' aspects of the work for others to pick up and work on themselves. These general properties of coordinate action are manifest concretely in situated practices that create and sustain a 'common field of work', whether coordinating 'walking' in the company of others or the 'sharing' of personal data with processors.

The common field of work in HDI is the catalogue of data sources that users generate. Data 'sharing' is organised around the catalogue and is ostensibly coordinated through the interactional arrangement request-permission-audit. This is an insufficient arrangement when seen from the perspective of cooperative work, however, for reasons that Schmidt points out.

> " … in order to be able to conceptualise and specify the support requirements of cooperative work we need to make a fundamental analytical distinction between (a) cooperative work activities in relation to the state of the field of work and mediated by changes to the state of the field of work, and (b) activities that arise from the fact that the work requires and involves multiple agents whose individual activities need to be coordinated, scheduled, meshed, integrated, etc. — in short: *articulated*. (ibid.)

Requests, permissions and audit logs are mechanisms of coordination within the field of work itself, but they do not articulate the field of work. They *order the flow* of information between users and third parties, but the flow itself stands in need of articulation. What, for example, occasions a request being made and being made in such a way for it to seem 'reasonable' to a user? Consider the expectations we might ordinarily entertain and the potential responses that might attach to requests from strangers, for example. Add to the mix how we might ordinarily react to requests regarding our personal data from strangers and it soon becomes clear that making a request is a non-trivial matter; that it requires *articulation*. As Bannon and Schmidt remind us,

> "Building computer systems where work is seen as simply being concerned with 'information flow,' and neglecting the articulation work needed to make the 'flow' possible, can lead to serious problems." (Schmidt and Bannon 1992)

Thus, a key design challenge in HDI is not only one of developing appropriate mechanisms to coordinate the flow of information within the field of work, but of articulating and thus coordinating *the work that makes flow possible* as well.

What does this entail? Schmidt (1994) highlights several generic features of 'social mechanisms of interaction' to support articulation work – 'salient dimensions' of cooperative work arrangements, such as who, what, where, when, how, etc. Schmidt suggests that these salient dimensions constitute 'elemental objects' implicated in the articulation of cooperative work arrangements (in contrast to the field of work itself) and that they provide a conceptual foundation for constructing computational mechanisms of interaction that support articulation work. Their elaboration geos beyond the 'minimal' representations of purpose wrapped up in requests in HDI to include *actors* (e.g., the particular parties involved in data processing); *roles* (e.g., the responsibilities that the particular parties involved processing data have); *activities* (e.g., the sequence of discrete 'jobs' implicated in processing the data and their status); *tasks* (the specific jobs being performed and their outputs).

There is more to Schmidt's elaboration of salient features of articulation work, and whether or not they constitute an adequate stipulation for articulation work in HDI or not is besides the point. The point is that no such stipulation currently exists in HDI. Neither the request or audit function provide adequate support and with it insight into the cooperative arrangement of work between users and third parties or the status of data processing *within that arrangement*. Cooperative work in HDI effectively occurs within a black box. A user cannot tell then from either the request or the audit such things as where in the arrangement of work the processing of data has reached, who is doing what with it, what's going to happen next, if there are problems or issues of concern, and so on. The articulation of work is limited to who wants the data for what purposes and reviewing such information. There is then very little support within HDI as it stands for the *ongoing management of relationships* between the various actors implicated in personal data sharing. Again, it is hard to see on what basis HDI could become a stable socio-technical infrastructure in everyday life without such mechanisms.

A key challenge thus becomes one of creating computational mechanisms of interaction that build the 'elemental objects' of articulation work into HDI to *make* 'salient dimensions' of distributed action *accountable* to users, thereby enabling them to manage and coordinate interaction. In saying this, we are not saying that we should blindly follow prior stipulations of salient features (though it does seem that some will hold), but that we need to develop a much better understanding of *what needs to be articulated* with respect to personal data sharing and the cooperative work arrangements implicated in it.

The same applies to the field of work itself. Schmidt points out that the distributed activities of a cooperative work arrangement are articulated *with respect to* objects within the field of work itself (e.g., data sources within the catalogue). A key issue here revolves around the 'conceptual structures and resources' that order the field of work, which enable members of a cooperative ensemble to make sense of it and act upon it. Again the question of interactional adequacy arises when we ask what conceptual structures HDI provides? It's not that it doesn't provide any, but the terms in which it does so are problematic from an interactional perspective. Take, for example, the dataware catalogue. It is conceptually ordered in terms of 'tables' that render data sources intelligible in terms of accounts, applications, installs, and services, etc. The problem in this is that the conceptual structure of HDI as instantiated in dataware is rendered in terms of the underlying technology, rather than in terms of what is being done through that technology, such as the processing of biological data as part of a healthcare regime. The problem thus involves ordering the field of work such that it *reflects* the work-being-done, or the work-to-be-done, rather than the underlying technical components of that work. It is hard to see then how users can articulate their distributed activities with respect to objects in the field of work when those objects (data sources) lack legibility or intelligibility to the broader populace in contrast to computer scientists and software engineers. Other, more 'user friendly' (and more pointedly) data-relevant, service-specific conceptual structures and resources are required.[3]

Gaining Traction: Interactional Challenges for HDI

Before we address the interactional challenges that confront HDI it is worth reviewing the problems that occasion them. We have seen in our treatment of personal data as a boundary object that data is not an object-in-itself, but an object possessed of various accountable social characteristics or 'dimensions', which ultimately embed it in mundane infrastructures. We have seen from a socio-technical perspective on infrastructure that human relationships are essential to the production and use of data, and that these relationships turn upon standardised methods of communication and coordination, which embed infrastructure in the organised practice of everyday life. We have seen too that mutual dependence is built into data sharing and that this occasions articulation work and the need to

[3] The requirement is reflected in the Article 29 Data Protection Working Party report on the IoT (14/EN WP 223, 2014) and the recommendation that end-users be able to "locally read, edit and modify the data before they are transferred to any data controller … Therefore, device manufacturers should provide a user-friendly interface for users who want to obtain both aggregated data and/or raw data." The challenge, of course, is bring this about in practice, particularly as personal data sources expand and diversify with the advent of the IoT.

build computational mechanisms of interaction to support it. Our purpose in reviewing salient work in social studies of science and CSCW has not been to define what a boundary object, infrastructure or articulation work is (by which measure this paper will no doubt be found wanting). Rather, our intention in selectively invoking certain features of salient texts has been to make it perspicuous how HDI becomes problematic when seen through a social or collaborative lens. Thus, we can see now that two key challenges confront HDI: one revolves around articulating the field of work in HDI, the other around articulating the cooperative arrangements of work implicated in HDI. We treat each in turn below

Articulating the Field of Work in HDI

In working our way through social studies of science and foundational CSCW texts we have seen how they occasion particular kinds of problem for HDI. We have seen that a mutual modus operandi is not in place and that the user whose data is being purposed by others does not have reciprocal opportunities for discovery. We have seen that data is not only 'mine' but 'ours' and thus social in character. We have seen that ownership and control are not isomorphic and that the life world drives the dynamics of these aspects of interaction. We have seen that data sharing is recipient designed. And we have seen that the conceptual structures and resources ordering the field of work lack legibility, intelligibility, and accountability in short. Each of the problems we have picked up on during our historical journey is an inherent feature of the field of work in HDI and presents challenges to its ongoing articulation.

User-driven Discovery

There are various aspects to the 'discoverability' problem, though of particular issue is what exactly should be made discoverable, and what kinds of control can users exercise over the process of discovery? These issues prospectively turn upon the articulation of *metadata* about a user's personal data sources, ranging (for example) from nothing more than articulating where a user's catalogue or catalogues can be contacted to more detailed information concerning catalogue contents. The demands of articulation work place further requirements on this process however, for even if users are willing to *publish* metadata about their data some means of understanding who is interested in discovering it may well be needed to build trust into the process – e.g., providing rich *analytics* into which processors are interested, when, how often, etc. Such analytics might provide users with resources that enable them to decide what of their data to expose or hide, though discovery may also turn in important respects upon other aspects of access control (e.g., defining pre-specified policies on who can and can't discover

their data).[4]

The issue of how users might drive the discovery process (finding data processors for themselves, whether for personal, financial or social purposes) is, however, more problematic and not something that has been addressed within HDI to date. Nonetheless, we would suggest that the discovery of data processors might be much like discovering new apps, and that the *'app store' model* may be a promising one to explore. Users are familiar with and make a conscious choice to visit app stores, where they are provided with rich metadata about apps and app authors that shapes their decision-making. Not only could data processors be 'vetted', much like apps in the iTunes Store, and detailed information about processing be provided, much like app 'permissions' in the Google Play Store, the social aspects of app stores also play an important role in the discovery process. User ratings and social networking links are important ingredients in the mix and help build the trust between users and service providers that is essential in the discovery and adoption of new technologies.

From My Data to Our Data

It is clear that the individuated model of ownership and control is not sufficient for real world applications of HDI. The social challenges of data ownership and control make it necessary to consider how individual and collective data sources can be collated and collaboratively managed by users. Individuals will not only need resources that enable them to control their own personal data sources, but will also need resources that allow them to *delegate* control of data sources and catalogues to others such that (for example) 'I' can assign control of 'my' data sources to 'you'. How ownership and control relationships are represented within and between catalogues, and what mechanisms will be needed to provide adequate support for their ongoing articulation, is an open matter, though transparency/awareness will be an important matter to consider along with rights management.

The creation and curation of collective data sources is an equally challenging matter. In one sense this may appear trivial. We can readily imagine, for example, that energy consumption data might relate as it does now to the household rather than specific individuals and that no complex identity and management issues are involved in such circumstances. Purposing such data is anything but a trivial matter, however. Who has the right to view and share such data? Who can edit it or revoke its use? Who actually owns and controls it? One view might be to default to the bill payer, but not all collective data sources are necessarily premised on contractual relationships. Add to the mix a world in which personal

4 All of this, as with so many interactions within the dataware model, trades on reliable identity mechanisms. The general problem of authentication in networked systems has been long studied and several solutions exist: TLS certificates (both server and client) or PGP-based web-of-trust seem feasible initial approaches, though both have weaknesses and would require careful engineering with respect to HDI.

data harvesting becomes increasingly associated with the things that we mundanely interact with, and the possibility of opening up both collective and individual behaviours to unprecedented scrutiny through data analytics becomes a real and problematic prospect. The inherent tension between individual and collective data will require the development of *group management* mechanisms that support *negotiated* data collection, analysis and sharing amongst a cohort.

The Legibility of Data Sources

Both the individual and negotiated production, analysis and sharing of personal data turn upon data sources being legible or intelligible to users. If users are to have the ability to exercise agency within an HDI system in any meaningful way, data sources must provide a minimum level of legibility as to what data they contain, what inferences might be drawn from that data, how that data can be linked to other data, and so on. Without some means to present this critical information, preferably in some form that can be standardised, it will be difficult for users to even begin to understand the implications of decisions they may make and permissions the give for processing of their data. As part of this it is key that users are not only able to *visualise* and inspect the data held by a source, but that they can also visualise and thus understand just what a data processor wants to take from a source or collection of sources and why – that just what is being 'shared' is transparently accountable to users, which may also involve making external data sources (e.g., consumer trends data) visible so that users understand just what is being handed over. Coupled to this is the need to enable *recipient design* by users. There are two distinct aspects to this. One revolves around enabling users to *edit* data, redacting aspects of the data they do not wish to make available to others both within a cohort and outside of it. The other revolves around *controlling the presentation of data* to processors when the accuracy of data needs to be guaranteed (e.g., energy consumption readings).[5]

Articulating Cooperative Arrangements of Work in HDI

Our selective trawl through the past has also made it perspicuous that HDI provides limited support to a key area of interaction: the articulation of cooperative arrangements of work implicated in personal data harvesting. This, in turn, raises the need to develop computational mechanisms of interaction that *surface* and *make visible* 'salient dimensions' of the cooperative work arrangements implicated in HDI to users. This goes beyond the interactional arrangement of request-permission-audit that orders the flow of information within the field of work itself to focus attention on enabling parties to the work to

[5] Controlling presentation of your meter readings may seem odd, but in a near future world where metering could be done on an appliance or device level, enabling users to control the granularity of energy consumption data (for example) becomes a much more coherent proposition.

manage the flow of information between them, including data interactions between internal members of a cohort and not only external parties.

Salient Dimensions of Collaboration in HDI

While it is clear that users will need to know who wants their data and for what purposes, our reflections have suggested that there is more to the articulation of data sharing than that. Requirements here are also admittedly vague – just what will users need to know about the cooperative arrangement of work in HDI to make the process work? Understanding this issue is a core research challenge and while our understanding is vague at this point in time it is clear that HDI will need to move beyond retrospective interrogation of audit logs to *real time* articulations that reflect the data sharing process itself. We might expect that the processing of data sources is an ongoing matter (as, for example, in the case of energy monitoring) and that this is something that users may want to monitor. Understanding the amassed body of outputs of ongoing data processing and the implications of this is something that users may well be interested in too. Ditto subsequent processing that might be applied by data consumers (e.g., the aggregation of personal data into big data sets). It is also clear that data consumers pass personal data on to third parties. *Tracking* what is being done with 'my data' and/or 'our data' becomes an important matter to consider then, articulating the *treatment* of personal data by data consumers, along with the development of mechanisms of that support this (e.g., preserving the provenance of data to enable tracking, notifying users of data reuse and transfer, and opening up such events to inspection and intervention).

The Incomplete and Open Status of Articulation Challenges in HDI

The challenges of articulating personal data within HDI are not settled matters. Rather, they open a number of *thematic* areas for further investigation, elaboration and support:

- ***Personal data discovery***, including meta-data publication, consumer analytics, discoverability policies, identity mechanisms, and app store models supporting discovery of data processers.
- ***Personal data ownership and control***, including group management of data sources, negotiation, delegation and transparency/awareness mechanisms, and rights management.
- ***Personal data legibility***, including visualisation of what processors would take from data sources and visualisations that help users make sense of data usage, and recipient design to support data editing and data presentation.
- ***Personal data tracking***, including real time articulation of data sharing processes (e.g., current status reports and aggregated outputs), and data tracking (e.g., subsequent consumer processing or data transfer).

Each of these themes stand in need of *interdisciplinary* investigation and elaboration, including ethnographic studies of current practices of individuals and groups around personal data creation and curation, co-designed interventions to understand future possibilities, and the engineering of appropriate models, tools and techniques to deliver the required technologies to support the complex processes involved in HDI and mesh the articulation of personal data with the organised practices of everyday life. What this amounts to in many respects is a call to the broader CSCW community to engage with the study and design of boring things – infrastructures – for personal data is embedded within them: in health infrastructures, communication infrastructures, financial infrastructures, consumption infrastructures, energy infrastructures, media infrastructures, etc. It is a call to study and build HDI around the unremarkable ways in which personal data is produced and used within the manifold infrastructures of everyday life, so that we might understand how personal data is accountably traded within human relationships and thereby develop actionable insights into what is involved in articulating those relationships in the future.

Conclusion

This paper set out to understand how interaction is configured within the field of Human Data Interaction, taking the Dataware infrastructure as an exemplar, and how this 'fits' with existing social viewpoints on personal data interaction. Seen from a social perspective, data interaction appears to be as much about *human relationships* as it is about data itself. Data, as Star makes visible, is always embedded in human relationships, and efforts to create infrastructure turn upon stabilising those relationships through appropriate methods of communication and coordination. CSCW orients us to key issues involved in creating such methods, particularly the need to devise mechanisms of interaction that *articulate* a) the field of work and flow of information between parties, and b) the arrangements of collaboration that make the flow possible. Historical insights drawn from social studies of science and CSCW have allowed us to identify a range of problems that affect HDI and a number of distinct thematic challenges they occasion. The broad challenge now is to address these problems and themes and shape the articulation of HDI around the accountable social nature of personal data interaction in order to drive a real and significant step-change in everyday life.

Acknowledgement

The research on which this article is based was funded by RCUK research grants EP/M001636/1 & and EP/K003569/1.

References

14/EN WP223 (2014) "Opinion 8/2014 on recent developments on the internet of things", *Article 29 Data Protection Working Party*, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp223_en.pdf

Bannon, L. and Bødker, S. (1997) "Constructing common information spaces", *Proceedings of ECSCW '97,* pp. 81-96, Lancaster, Kluwer.

Cafaro, F. (2012) "Using embodied allegories to design gesture suites for human-data interaction", *Proceedings of UbiComp '12*, pp. 560-563, Pittsburgh, ACM.

Gerson, E. and Star, S.L. (1986) "Analyzing due process in the workplace" *ACM Transactions on Office Information Systems*, vol. 4 (3), pp. 257-270.

Haddadi, H., Mortier, R., MaAuley, D. and Crowcroft, J. (2013) *Human Data Interaction*, Cambridge Computer Laboratory. www.cl.cam.ac.uk/techreports/UCAM-CL-TR-837.pdf

Haddadi, H. et al. (2015) "Personal data: thinking inside the box", *Computing Research Repository*, http://arxiv.org/abs/1501.04737

Kee, K., Browning, L., Ballard, D. and Cicchini, E. (2012) "Sociomaterial processes … towards effective collaboration and collaboration tools for visual and data analytics", *Science of Interaction for Data and Visual Analytics Workshop*, Austin, NSF.

McAuley, D., Mortier, R. and Goulding, J. (2011) "The dataware manifesto", *Proceedings of the 3rd International Conference on Communication Systems and Networks*, pp. 1-6, Bangalore, IEEE.

Mashhadi, A., Kawsar, F. and Acer, U. (unpub. manu.) *Human Data Interaction in the IoT*, Bell Laboratories. www.fahim-kawsar.net/papers/Mashhadi.WF-IoT2014-Camera.pdf

Mortier, R., Haddadi, H., Henderson, T., McAuley, D. and Crowcroft, J. (2013) "Challenges and opportunities in human-data interaction", Proceedings of the 4th Digital Economy All Hands Meeting, Salford: RCUK. http://de2013.org/wp-content/uploads/2013/09/de13-hdi.pdf

Mortier, R., Haddadi, H., Henderson, T., McAuley, D. and Crowcroft, J. (2014) "Human-data interaction: the human face of the data-driven society", *Social Science Research Network*, http://dx.doi.org/10.2139/ssrn.2508051

Pentland, A. (2012) "Reinventing society in the wake of Big Data", *Edge*, 30-08-2012. http://edge.org/conversation/reinventing-society-in-the-wake-of-big-data

Schmidt, K. and Bannon, L. (1992) "Taking CSCW seriously", *Computer Supported Cooperative Work: The Journal of Collaborative Computing*, vol. 1 (1), pp. 7-40.

Schmidt, K. (1994) *COMIC Deliverable 3.2 'Social Mechanisms of Interaction'*, Esprit Basic Research Action 6225, ISBN 0-901800-55-4.

Star, S. L. and Griesemer, J. (1989) "Institutional ecology, 'translations' and boundary objects'", *Social Studies of Science*, vol. 19 (3), pp. 387-420.

Star, S. L. (1999) "The ethnography of infrastructure", *American Behavioral Scientist*, vol. 43 (3), pp. 377-391.

Star, S. L. (2010) "This is not a boundary object: reflections on the origin of a concept", *Science, Technology and Human Values*, vol. 35 (5), pp. 601-617.

Strauss, A. (1985) "Work and the division of labor", *The Sociological Quarterly*, vol. 26 (1), pp. 1-19.

Suchman, L. (1995) "Making work visible", *Communications of the ACM*, vol. 38 (9), pp. 56-64.

Tolmie, P., Crabtree, A., Rodden, T., Greenhalgh, C. and Benford, S. (2007) "Making the home network at home: digital housekeeping", *Proceedings of ECSCW '07*, pp. 331–350, Limerick, Springer.