Distributed Ontology Building as Practical Work

Dave Randall\textsuperscript{2}, Rob Procter\textsuperscript{1}, Yuwei Lin\textsuperscript{3}, Meik Poschen\textsuperscript{1}, Wes Sharrock\textsuperscript{1} and Robert Stevens\textsuperscript{1}
\textsuperscript{1}University of Manchester
\textsuperscript{2}Manchester Metropolitan University
\textsuperscript{3}University of Salford

Abstract

Ontologies - a form of structured and logically related knowledge or classification hierarchy embedded in a computer system - are regarded by many scientists as having enormous promise for the consistent use and re-use of data. To realise this promise, however, is not straightforward. In this paper, based on ethnographic observation, we argue that the challenges for ontology building are 'social' as much as they are technical. By this we mean, the routine work undertaken in the building process and the problems and difficulties entailed can be understood in terms of the practices of knowledge workers and the practical nature of 'sorting things out'. Getting a better sense of how, in practice, this work gets done gives a sense of the main challenges of building successful ontologies and how this impacts on the design of tool support. In considering the practices of one group in particular, we try to show how, for members, the technical problems of determining what classification structure is appropriate, and what its boundaries might be, depend substantially on assumptions about the 'community' and its interests and purposes. This 'turn to the social' has ramifications for the understanding of ontology building and use. Specifically, 'modelling' approaches to ontology building tell us little about the practical organisation of the work and how this relates to the prospect of successful sharing. Ethnographic enquiry may reveal important issues that are otherwise missed.

1. Introduction

Ontologies provide a means to formalise knowledge in machine-processable forms. Formalisations of this kind can be subjected to machine 'reasoning' which reveal the full set of logical relationships between various instances and classes, and also where logical inconsistencies are to be found. Recent interest in ontologies can be traced to the Semantic Web and its vision of Web users being able to have 'intelligent agents' assist them in discovering and interpreting information on the Web (Berners-Lee \textit{et al.}, 2001). More recently, this interest has been further fuelled by the Semantic Grid, and its goal of automating the discovery and composition of distributed information resources and processing services (De Roure \textit{et al.}, 2001). The application of ontologies to knowledge representation in scientific research has been a particularly active area\textsuperscript{1}, as researchers struggle with the challenge of making more effective use of increasingly vast amounts of data and information (Hey and Trefethen, 2003). As a means to codify data in consistent, structured, ways they hold enormous promise for the use and re-use of scientific data.

\textsuperscript{1}See http://www.w3.org/TR/webont-req/
Although by no means universally accepted, ontologies are in common use in scientific domains such as biology. A significant part of the value of an ontology derives from it being shared and so we may think of an ontology as a form of cooperative system or infrastructure. It follows, however, that the degree to which an ontology is adopted may depend both on technical adequacy and on the degree to which community interests, purposes and politics can be managed— we gloss the latter as being ‘social’ concerns. As we shall see, the technical and the ‘social’ turn out to be closely intertwined.

The aim of this paper is to examine, mainly through a single case study, how a group of cell biologists and bio-informaticians go about constructing an ontology; the kinds of routine issues that arise for them, and what the implications might be for ontology-building and ontology use. In fact, we will argue that what makes this especially interesting is firstly that ontology building is a form of knowledge work that entails relatively little separation between ‘designer’ and ‘user’ and, secondly, that technical problems closely associate with issues concerning who should be involved in the building and what community of users is imagined. A significant part of the value of an ontology derives from it being shared, so we may think of an ontology as a form of cooperative system or infrastructure. As Bowker and Star (1999) observe, ‘creating infrastructure is as much social, political and economic work as it is theoretical.’ (p. 109). We might expect, then, that one of the main difficulties in successful ontology building is arriving at a consensus among builders and users and this is borne out by our findings. Finding that consensus, however, is not easy. In the projects we have studied, much time and effort is spent reaching agreement about what should be in a given ontology and what should be left out. This, in turn, raises questions concerning what is the ‘right way’ to go about building an ontology, what tools should be used in its construction and how to support the process (Corcho, Fernandez-Lopez and Gomez-Perez, 2003).

The point we will make is that decisions, policies and strategies for ontology building can be thought of sociologically as issues to do with the assumptions that knowledge workers make when they do their work, the practical nature of their enquiries and the beliefs they carry when they make them. The specific variety of sociology in question is what has been termed ‘ethnomethodologically informed ethnography’ (see Randall et al, 2007). There is no space here to detail the analytic commitments entailed in this view, but it entails a rigorous commitment to detail, to understanding whatever is being done as meaningful and interactional ‘work’, and a refusal to engage in the epistemological and ontological questions that more conventional sociologies are concerned with. As such, it concerns itself with, and only with, the way in which ‘members’ - skilled and competent practitioners - accomplish the tasks they set themselves. It is, to put it succinctly, a practical sociology.

Ethnomethodologically-informed ethnography has been regularly deployed in the context of Computer-Supported Cooperative Work (CSCW) (see e.g. Randall et al, ibid) and Science and Technology Studies (STS) (see Lynch, 1997). In relation to the former, it has been used very effectively to challenge top-down modelling processes in requirements engineering (see Goguen and Jiroinka, 1994). The claim was that models of this kind should be complemented by a recognition of complexities that incorporate the ‘social’ (Anderson, 1994; Hughes et al., 1993) and that ethnography can be make an effective contribution to requirements gathering and design decision-making. STS has also used
ethnographic techniques, although in a rather more evaluative way - and to rather more theoretical purposes - in the field of e-Science (see e.g. Baker and Ribes, 2007; Baker and Millerand, 2007).

There have been some recent attempts to incorporate this sociological notion into classification work in general and ontology-related work more specifically (see Pike and Gahegan, 2007). We wish to develop this. Our ethnographic approach provides, we suggest, some purchase on the practical issues relating to ontology building; an analysis which orients to the context-specific, and a preliminary analytic framework with which to understand the process better (see e.g. Ure et al., 2007; Lin et al., 2007; Lin et al., 2008).

In recent years, ethnographers have turned to more complex domains and to the knowledge work associated with them (see e.g. Harper et al., 2000). Here, the emphasis has been on a conception of knowledge that emphasises the practice of knowledge production. The dominant metaphor in the sociological approach to classification has become that of ‘expertise sharing’ (Ackerman et al., 2003) and research has emphasised the ordinary, practical ways in which knowledge or expertise is (or is not) shared across knowledge communities.

This involves thinking of ontology or classification building as ‘work’. Such a view owes much to Bowker and Star (1999), who point out that our understanding of classification systems often only report on the final result – the classification scheme itself – and does not incorporate any appreciation of the work that went into its construction. Our purpose is to identify how this problem of knowledge work and its behavioural dimensions resonates with a number of issues, which in Bowker and Star's terms are ‘challenges’ for classification schema in terms of comparability, visibility and control.

Comparability refers to ‘regularity in semantics and objects’ (ibid: p. 231) which refers to the problem of whether and to what degree terms can be defined in consistent ways and hence shared unproblematically. We will endeavour to show that this comparability is not easily arrived at in the context of ontology building and is the source of regular revisiting, even in relatively homogeneous communities of users where one might assume underlying concepts (if not terminology) stand a good chance of being commonly held. Ontologies that should serve more heterogeneous purposes may turn out to be serving one user group more successfully than another – a problem that has been well-attested to in the field of medical informatics (Rector, 1999). As we show below, achieving this regularity entails the use of a wide range of resources including both artefacts and the development environment.

The second of the challenges to classification recognised by Bowker and Star is visibility. ‘Invisible’ areas of work are ‘by definition unclassifiable except as the residual category: “other”’ (ibid: p. 231) ‘Invisible’ work can refer to those informal practices which do not, in themselves, constitute part of the ontology, but which may, nonetheless, be critical to how an ontology is constructed. As we shall see, this has ramifications, both in terms of decisions concerning what shape the ontology should take (the axes of classification), but also defining method and scope of the ontology in question.

The third challenge is that of control which, in Bowker and Star’s terms, is a measure of the degree of prescription that a classification scheme imposes on its users. In the context
of ontology building, we argue that this manifests itself largely as a problem of ‘enlistment’, entitlement and purpose – who should be involved in the building of ontologies, how and when. As we shall see, this is not trivial.

If Bowker and Star are right in that these issues are central to any classification system, and, if they pertain as much to ontology-building as to any other kind of classification scheme, then it would seem that there are good reasons for examining the work that goes into the construction of ontologies, because that work will ramify in the development of ontologies and of tools to support it. Their success or failure when deployed will not only be a matter of their internal consistency, but also the degree to which they meet organizational requirements. There may be a number of dimensions to this and below we sketch out what some of them may be, based on our own observations of ontology-building work conducted over a period of more than a year.

2. The Case Study

The evidence we reproduce below comes from an ongoing study of work by teams of people engaged in the building, maintenance and use of Protégé-OWL, a standard ontology building tool, and groups involved in developing and promoting the use of ontologies for scientific research. Over a period of more than 12 months, we have conducted a series of focus groups, ethnographic observation of face-to-face meetings and interviews with members of the Collaborative Open Ontology Development Environment (CO-ODE) project\(^2\) which aims to develop authoring tools and infrastructure that make ontology building easier; a number of ontology building projects, including NeuroPsyGrid, whose objective is to devise a common ontology for neurosciences; and the wider ontology user community, including members of the Ontogenesis Network\(^3\) a ‘network of excellence’ which exists to foster the creation and adoption of ontologies in the biological sciences. Interviews were recorded and transcribed and meetings videoed for subsequent analysis. We also had access to six months of emails circulated to members of the Ontogenesis Network mailing list.

2.1 An Overview of Ontologies, Languages and Tools

Ontologies consist of a set of classes representing the categories of the entities of interest in a domain and the relationships between those entities. In doing so, an ontology can be used to capture ‘what it means to be one of those entities; that is, the semantics of the domain’. Ontologies are described using a formal language that provides:

1. A single uniform, concept classification scheme that provides a controlled vocabulary (see Figure 1).
2. A means to define the properties of each concept as a series of attributes (see Figure 2).
3. A means to define a concept’s relationship with other concepts, usually in the form of a hierarchy (see Figure 1).

\(^2\) [http://www.co-ode.org/](http://www.co-ode.org/)

\(^3\) [http://www.ontonet.org/](http://www.ontonet.org/)
In the figures below we see extracts from an ontology for disease involving three high level concepts: ‘Disease’, ‘Organ’ and ‘Organism’. In Figure 1, we see the top level components of the concept hierarchy. In Figure 2, we see that ‘Disease’ has two attributes: ‘has_focus’ and ‘has cause’, and in Figures 3 and 4, we see how these are used to specify relationships that ‘Disease’ has with ‘Organ’ and ‘Organism’. Figures 5 and 6 show how ‘Pneumonia’ inherits attributes from ‘Disease’ (its superclass). Figure 7 illustrates how, using the ontology, a reasoner is able to infer that ‘Bacterial Pneumonia’ is a type of both ‘Lung Disease’ and ‘Infectious Disease’, even though the latter is not explicitly represented in the ontology (Kola et al. 2010, reproduced by permission).

Figure 1: A conceptual hierarchy.

Figure 2: Disease concept attributes.

Figure 3: The description for ‘Lung Disease’.
Figure 4: The description for ‘Infectious Disease’.

Figure 5: The description for ‘Pneumonia’.

Figure 6: The description for ‘Bacterial Pneumonia’.

Figure 7: The inferred conceptual hierarchy for ‘Bacterial Pneumonia’.
The projects in our study employ a number of tools for ontology development. These include: the Web Ontology Language (OWL)\(^4\), a commonly used language for expressing ontologies, and Protégé-OWL, a leading ontology engineering environment, a suite of inter-operable tools that includes ontology editors, visualisers; checkers (for testing correctness); ‘reasoners’ (i.e. software tools that can derive implications from a set of facts or axioms); version management and collaboration support.\(^5\)

3. Ontology Building as Expertise Sharing

The problem of attaining stability in a classification scheme (one that is strongly implicated in any model of ontology maturation) can be understood, as we have indicated, as a practical problem for members; one that depends on the kind of obstacles that typically crop up.

3.1 Control - Enlistment, entitlement and purpose

The first and most obvious obstacle, as already suggested, is that communities of users are becoming more heterogeneous in their interests and practices. The notion of ‘community’ crops up again and again in our data on ontology building work. This should not astonish us, for some part of that work is the business of imagining, or actually finding out, who might be involved in the building of, and the subsequent use of, the ontology. Nevertheless, this can be a difficult matter. Firstly, there is an issue of entitlement – who should be invited to be involved in the process. Below, we show how a variety of factors play a part in the way participants get involved. In turn, as we shall see, this affects how ontologies evolve.

Unsurprisingly, perhaps, it sometimes proves difficult to enlist members of the community. This relates to what in CSCW is recognised as Grudin’s insight (Grudin, 1994) in respect of successful collaboration tools, which is that they require that those doing the work benefit in some way from it. As one ontology builder put it,

*The trouble is, it involves an awful lot of drudge work, especially in respect of coming up with definitions that everyone agrees with…*

It became apparent that for our ontology builders a significant problem was not only who to get involved, but when to involve them. Much of our data concerns the way in which it is difficult to get user community involvement at the point where the builder needs it, but easy when others in the community see it in terms of their purposes. We would suggest, however, that where ontologies are to be deployed in complex environments, a great deal more exploration of what those purposes might be is needed. The reasons for this quickly become obvious, as one participant, engaged in a collaborative process of building an environmental ontology pointed out:

*We’re going to have to consult a lot over terms, but maybe will have to legislate to some degree. We’re going to have to allow people to tag stuff up. To some extent, at least to begin with, there are some agreed boundaries… tropical rainforests have agreed attributes… and these would be largely unchallenged. There are, however, about twenty different types of grassland within that habitat, defined not only by the*

\(^4\) http://www.w3.org/TR/owl-ref/  
\(^5\) http://protege.stanford.edu/
type of grass, but also maybe by the fauna as well. And then there’s all this to do with
temperature, rainfall, etc, etc. We wouldn’t want to include geospatial information.
Habitats grow and shrink, but we don’t need that...

The issue of ‘consultation over terms’ again has practical consequences for ontology
building, for there is controversy over when and how that should take place. As the above
quote indicates, some initial decisions about what ‘we’ might want or need are made
before the ‘we’ in question has been fully defined. On the one hand, as another
participant observed, ‘GO [Gene Ontology] started from a use case… that’s why it was
so successful…’ In other words, early user involvement can be seen as desirable. On the
other:

It’ll just be a controlled vocabulary to begin with… with most successful ontologies,
the complexity came later. Sometimes you feel like a lawyer, finding descriptions that
no-one’s going to disagree with...

Hence, there is an implication that the appropriateness of different building strategies
rather depends on the way in which the trade-off between the value of keeping it simple
in order to expedite the process (or at least get it under way) and the desire to achieve
sufficient complexity (and hence modelling power) is managed. Thus, ambition might be
limited to the construction of a folksonomy – i.e. a collaboratively produced, practice-
based and emergent classification scheme or taxonomy. It might entail a ‘lightweight’
tonology (e.g. taxonomy) that is relatively informal in the way it defines a domain, to a
‘heavyweight’ ontology where all aspects of the domain are formally defined. Again, it is
surprising how little is known about the different conditions under which folksonomies,
taxonomies, ‘heavyweight’ ontologies and so on might be appropriate. It seems to be a
common experience among ontology builders that getting any kind of community
agreement over terms in the earliest stages is extraordinarily difficult, perhaps because
common purpose is difficult to find. In one group we studied in some detail6, the task was
to ‘rebuild’ a Cell Type Ontology (CTO). People were invited to participate on the basis
of expertise in cell biology and in ontology building and, moreover, on the basis of
‘people we can trust... I know I can get on with them…’, as one participant remarked.
During the period we observed the group, a considerable amount of time was spent on
identifying what the interests of user communities might be and who was entitled to
represent them. This involved identifying what people had already contributed and, as it
were, their importance7:

P1: from what I understand is that the OBO8 people have commissioned a reworking
of the CTO... and I’m perfectly happy for this to be a contribution... but that is not
something I will manage... because the whole process would just drive me up the
wall...

6 The group consisted of eight biologists and bio-informaticians from both commercial
and academic backgrounds, and a number of different institutions. At least five would be
well known either as cell biologists or bio-informaticians.
7 In the following extracts, participants have been anonymised.
8 OBO refers to Open Biomedical Ontologies, of which there are a number. GO, the Gene
Ontology and PATO, the phenotype ontology, referred to in this paper, are two of them.
See http://obofoundry.org
P2: If I could comment here... we’ve been using the CTO... we started to look at the hierarchy, but the fact that lots of things are not defined, they know there are lots of missing ‘is a’ relationships that need addressing... They had a discussion about rebuilding the whole thing again from scratch... start again...

P3: who is ‘they’ here?

P4: active are... [a list of names]... the CTO doesn’t have like a paid person to look after it... originally it was [other names] and now it’s just sitting there in no man’s land...

P3: but that no man’s land is located over in [location]...

P2: No not particularly, though most of those people are over in the [location]... X is in [location]... right now, Y is in [location] too... I don’t know where Z is...

The significance of such discussions is that understanding the nature of the existing community’s commitments has implications for the group’s understanding of its own purposes. Nevertheless, judgements have to be made about the community in question. It is noticeable that some part of what is being identified above is the ‘trustability’ of other work being done and the people doing it. It is quite evident that some part of the work that the group does initially is to arrive at some view of what its purpose might be in and through an assessment of the ‘state of play’ in various places. Equally, the process is defined in part by the interests of the individuals participating and the organisations they might represent. Hence what follows shortly after:

P2: I would like to use the CTO for my own uses, one of which is tying all the available public cell lines that we have data on and getting a type for them and making something cross-product which is something that really needs to be done... that’s somewhere on our list of what we need...

Such matters impact directly on what kind of ontology will be built. Its scope, size and boundaries relate directly to the use for which it is intended. These varying interests have all sorts of consequences for the way in which any proposed ontology might be structured. Moreover, and it is important for our assumptions concerning distributed work, this group both spent a great deal of time in defining its purposes, the relevance of previous work and the people engaged in it (in conjunction with the methodological work discussed below, approximately one and a half days of discussion). Further, this was one of the things that appeared to necessitate the whole group (some ten people) working together. In contrast, and as we shall explain below, the actual business of populating the ontology (i.e. creating instances of entities it defines) was done without this kind of work.

3.2 Visibility and Comparability - Defining method and scope

One feature of the work was the changing ambition and scope as the process developed over time and we can conceive of these issues as being to do with visibility. The original CTO had been built over a short period of time and ‘hardly touched since’. It was built by a small number of people in a couple of days. Hence, early in the meeting we hear:

P1: I want to take the OBO CTO... which is a hand-crafted taxonomy... it’s a multiple hierarchy... what X has described as a tangle.... what tends to happen when you build ontologies by hand is that you make mistakes... what we have discovered is that one in ten of the classes has a missing or erroneous subsumption relationship on it and the process of normalisation is supposed to give you reusable modules... more
maintainable lumps of hierarchy... and highly axiomatised ontologies... with more stuff in them... so that you can get more computational inferences... essentially it does all the work.

A number of issues are being claimed as relevant here and they implicate a notion of what a ‘good’ ontology might be. Firstly, what exists is a ‘tangle’ that implies several different axes of classification. Secondly, what exists contains a number of errors and, thirdly, it is only a ‘taxonomy’ and thus does not contain enough in the way of subsumption hierarchies to do what ontologies in this view are here to do – enable the derivation of computational inferences. The solution proffered is ‘normalisation’. Hence, ‘By the end of this couple of days I would like to come out with at least a plan for how we normalise it...’ Although not stated explicitly at this point, there is the implication that the meeting will endeavour to at least begin the process of developing something which can and will be used by a ‘community’ of people.

If the existing CTO is a ‘tangle’ and hence little ‘reasoning’ will be possible, then, a ‘good’ ontology needs to contain a clear hierarchy and, more particularly, one primary axis of classification if possible. What is entailed in this – to begin with – quickly becomes evident:

P1: The general flow of activity... we’re going to have a general look at the CTO... we need to have a look at the axes of classification... how are the cells classified... both explicit and implicit... you will see terms like... uh... up at the top you’ll see ‘cell by lineage’... what does lineage mean?... so that’s an explicit axis of classification... but you’ll also see terms like, ‘mature cell’ and ‘immature cell’, but that’s not an explicit axis... that’s just hidden within the names... the maturity of cells is something that we can pull out through a restriction... having identified the axes of classification we need to identify a primary axis of classification... all the other axes are then pulled out into supporting ontologies... a lot of these already exist in things like PATO... so the phenotype ontology, one of the axes is ploidy... and so we need to and have a look at how ploidy is described in PATO, and then we will be able to take the actual cells at the end of the leaves... And then for instance you can recreate the intermediate class... but it’s complete and it’s dynamic and it’s... lovely.

It is already obvious that interrogating other ontologies will form a significant part of this work, and also that the ‘leaves’, that is those cells that will eventually be at the bottom of the subsumption hierarchy, are not particularly important at this point. In other words, ‘completeness’ is not an immediate aim. The ontology can be more fully populated at a later stage. Those supporting ontologies that are the ‘right’ ones is fairly evident from the outset, but even so integrating this work with them will be a non-trivial process.

One of the surprising aspects of the process was that the business of finding this axis of classification was extremely time-consuming, required extensive discussion, and was ultimately rejected as the way forward. Put simply, a large part of the early business was to do with establishing what the right ‘plan’ for assembling the new version of the ontology might be. Again, very early on we see a candidate method put forward:

P1: What I’m hoping to do in identifying the primary axis is do this somewhat formally, using Ontoclean. What Ontoclean does... it’s a way of evaluating

---

9 A subsumption hierarchy is a classification of entities from the general to the specific.
subsumption relationships and checking that you've said the right things in the right way. It talks about unity, rigidity and identity... unity is all about whether you're talking about parts and wholes cos one of the common mistakes is to talk about part-whole relationships as 'is a' relationships. Famously, ocean is a kind of water where water is part of ocean... identity is all about necessity and sufficiency which I hope that, being OWL people, you're all reasonably familiar with. Rigidity is talking about things which are inherent to the... ummm... ah... ah... what properties are held by an entity for the duration of its existence or only part of its existence. And what we want to do or what we should do is identify the primary axis to be a rigid property... and helps us make a nice safe tree...

In effect, this means finding ways of describing cells that are always true for those types of cell (bearing in mind that cells can change over time). The attempt to delineate the function of the meeting was an important part of shaping the ontology itself through achieving a consensus about its scope and ambition. This has a number of features, the first of which is a general discussion about the possible scope:

P5: well, it's not that clear what the CTO was trying to model... a cell-type is an approximate synonym for class or concept... biologists distinguish between cell-types and cells so should this actually be the cell ontology...

P3: for use by biologists it would have to be the cell-type ontology.

P1: what I think one of the goals should be to make it so ontologically beautiful that it's unusable...

[general laughter]

There is more going on here than is immediately apparent. In effect, what is being said here has much to do with the difficult problem of who will use the ontology and, hence, for what purposes. It seems obvious that such an ontology would be used by biologists, but, in fact, different ontologies can cover the same broad domain, but model that domain very differently and be used in very different ways. Different communities of user can be and are envisaged. By way of example, Au et al. (2006) point to the way that three different ontologies, the FMA, GO, and the CCO\textsuperscript{10} all contain knowledge about cellular structure, but are nevertheless different. They show that an important part of this difference lies in purpose. As they say, ‘the FMA provides a framework for modelling generic anatomy, and thus, some higher-level terms are meant for mammals or vertebrates... In contrast, the GO models a canonical cell across multiple species and is designed for a specific purpose – annotation of genomic research and hence excludes particular cell types, unlike the FMA that models some cell types explicitly.’ The size, scope and ambition of this ontology, then, has to do with exactly who will use it.

The new version of the ontology at this point will be built from an axis that has to be agreed and the agreement will be arrived at by eliminating other candidates. Moreover, decisions about how much work is to be accomplished have yet to be made:

P5: one thing, above that level, are we just looking at cells in vivo? Or experimentally modified?

\textsuperscript{10} CCO is the Cell Component Ontology. See http://brg.ai.sri.com/CCO/
P1: They just have like six crosses... they say that this is very weak... they just haven’t populated it...

P2: It’s kind of horrible... I’d be tempted to put that somewhere else...
I was tempted to just look at cells in vivo, so basically we should just take that partition out...

P5: The only reason I was interested is because it’s non-canonical... most of the ontologies in OBO are canonical... so if you wanted to produce the full thing you essentially would make a cross product so you would take all of the terms in the one...
P2: so you’d compose it out of other things, but in this case the things that are being composed are part of the same ontology... so experimentally modified cells are something I’m particularly interested in, but I don’t think they belong here... this part of the CTO is really problematic and we shouldn’t go there...

P5: I agree with P1, it’s probably out of scope for this...
A full and complete CTO would have to include cells which are modified in laboratories. A decision evolves here not to include experimentally modified cells and is predicated on two elements. Firstly, that what currently exists is radically incomplete so there would be a considerable additional overhead and, secondly, that incorporating this category of cell would significantly increase the difficulty of the exercise. Later, we see a similar decision made in respect of single-celled organisms:

P3: let’s think about the purpose of this... if the purpose of this is to classify cell types in multi-cellular organisms... that’s what we should classify and forget the rest...
P1: ummm...
P3: we don’t need to classify cell types in yeast...
P1: we haven’t made that decision yet... ummm... we might have done [laughter]
P2: from a purely data point of view... 80% of our data is eukaryot (complex cells with membranes around them - most living things - nucleus inside the membrane) not prokaryot (mainly single cell organisms, no nucleus) and you do very different kinds of experiments with prokaryot... it’s almost never about cell type... and that’s something that’s universal across almost all the databases...

The proposed ontology, then, will only be populated with information about in vivo cells and cells to be found in multi-cellular organisms. What informs this decision is based on a combination of factors that relate to what is ‘doable’ and to the possible value of the results. In the latter instance, cells associated with single cell organisms are discounted largely on the basis that the kinds of experiment done with them do not require cell type information.

Space precludes a detailed examination of each candidate axis, but it turns out to be difficult, not least because establishing ‘reliable’ information in a situation where expertise is distributed unevenly is not straightforward. As we shall see, this socio-technical distribution of expertise is critical when we look at the information getting strategies of the group. A lengthy discussion ensues which involves the discussion of the various properties of cells that might be termed ‘rigid’. Each, in turn, is raised as a possibility, some are dismissed quickly and others treated as serious possibilities. It turns out that the rigidity of cell properties depends in large part on the way in which these properties are defined, no small matter when ontologies are being built. This fact leads to
one axis after another being rejected. The discussions result in the decision that no cell property is sufficiently rigid to form a primary axis and, therefore, a new approach to the building of the ontology will be necessary. This approach, by default, will be to create a list of cells under the heading ‘cell’, ascribe properties to them and assume that if that is done correctly the reasoner will sort the cells into a hierarchy:

P4: I think we may have got to the point where we cannot find a primitive axis...
P1: well, in that case we go for the ultra normalisation... of doing it all by restriction... so my current proposal is that we just have cell and we list all the actual cells underneath...
P5: so if we just have cell, are we making the assumption that everything in the CTO will hang under cell... so cell functions or processes would not be a type of cell, so we should have more than one upper level... we need classes as well as cells...
P3: we need types of function...
P5: we need a process hierarchy
P1: which, funny enough, we have in GO... so are we happy that we just have cell and do it all by restriction?
P5: well, not happy, but we haven’t found any property that we can treat as rigid...

It is only at this point, many hours after the group first met, that they begin the process of populating the ontology. They do this by selecting cells with which they wish to work. Again, one of the principles that the group orients to is that of tractability. Whatever decisions they make, the work has to be doable. It remains the case that, although completeness is mentioned as an issue, the group is happy to adopt a satisficing attitude, such that what they do will be ‘good enough’ for their purposes:

P5: so then our assumption would be that we put a load of cell types under cell and our hope would be that there will be none that can only be inferred.
P2: it’s just a question of completeness, isn’t it?... there are still things sitting there, it means we haven’t got properties we can find enough to build a good enough hierarchy, but actually it’s a more tractable problem... and actually we could do this by picking some sensible cell types representative of plants and animals; circulatory and secretory... it gives us a pretty good go at the restrictions...
P1: if we just go and pick twenty... and just do the restrictions... and then go back and generalise... what I propose now is that we assign some tasks that people can go and do... someone can go away and select twenty...
P2: we can do that collaboratively now...
P1: can someone write this down... one task is to select twenty or so of actual cells which give us a representative spread, one is to go away and find something that talks about morphology, process, nuclear number, most of these are going to be PATO by the way... ploidy, lineage we probably don’t need to bother with because its all there... and then there’s organism...

[P2. Notes them all down]
The ontology now will be built by applying restrictions to each cell type and these restrictions will either be defined by the group itself or will come from existing pieces of ontology they are able to grab. The group begins to use Protégé-OWL at this point, and a simple hierarchy with ‘cell’ at top and twenty five cells listed underneath appears on
screen in front of them (see Figure 8, left-hand pane; the right-hand panes display information about the selected cell class):

P1: So, what... we’ve now got twenty five candidate terms... next stage is to go and find bits of supporting ontology for dealing with the other axes of classification as identified this morning, as in function or process, taxonomy, morphology, staining, lineage, anatomy, but we’ll put anatomy to one side. Now pairs of us can look at these things... two pairs to look at PATO and the rest look at GO process... so what we need to do for PATO is whether the terms are there and then how they’ve done it... to see whether it actually has the classification that will give us what we need...

It might be assumed that this is merely a matter of marrying terminology from one ontology with that of another, but things are rather more complex. Firstly, and because ontologies have specific purposes, terms may be defined in different ways or may be incomplete:

P2: maybe you’d like to say what the problem is... and I think its that all the anatomy ontologies are species specific... and the CTO is not... and therefore we have to cross product many ontologies and not just one... which is a much harder problem...

P1: so its a watch this space problem...

P5: the other thing we’ve learned is that PATO is seriously short of synonyms...

P2: they’re definitely missing synonyms they know that

P3: PATO hasn’t been created with cells in mind... its animal oriented...

P2: its animal orientated because the people who are using it are using it to annotate mouse phenotypes...

P5: its very mouse driven... it doesn’t mean that’s the scope... it means that’s the practicality... they’d be very happy to make it more general... this is a very useful piece of work...

Secondly, it may turn out that the work requires some modification of an existing ontology that will in turn necessitate liaison with other members of the community:

(PATO is up on the screen) can we see it?... it’s got biological sex, then... but here’s nothing about chromosomal basis...

P2: so all of them come from the PATO ontology and those are in there by request of [reference to organisation]...

P3: male and female is all I would want... let’s get some procedure here, P2?. Will you undertake to request these changes to PATO...

P2: what I will do is undertake to produce a workshop record with some action items otherwise I’ll have to type them in one by one into the PATOtracker which isn’t something I want to do...

P3: but you can cause PATO to evolve

P2: what I can do is go and see [person B]... he’s in [place Q]... and then I’ll report back
Figure 8: Protégé-OWL user interface showing cell hierarchy.

It is only now that some attempt is made to actually construct an ontology, and even then only in a highly provisional form. Indeed, it is referred to as a ‘toy’.

P1: the other thing is, if in the toy we’ve got twenty five cell types... and we’re doing lineage... if we’ve got cell A and its predecessor is B then we’re going to have to put B in, and... in theory all the chain going back to zygote... but for the moment we’ll just fake it... so instead of having a b c d e d... we’ll go a b f g just so we’ve got a couple of steps so that the transitivity can be seen to be working...

The reason that a ‘toy’ version is in play here is evident. Useful work can be done here, developing a provisional hierarchy, ascribing properties to cells, without (again) having to worry about completeness and consistency. This will become more of a focus over the remainder of the period that the group is at work. The ‘toy’ functions as a means for everyone to see the cells that have been chosen, and as a means to begin the building process. After this, the group again divides such that pairs can undertake the work of establishing properties for each of the cells that have been included in the ‘toy’. Nevertheless, the ‘toy’ is not even the main focus of subsequent work. What happens is that copies of a spreadsheet containing a matrix of the selected cells and the properties that the group has decided to describe is developed, distributed among the pairs and taken away:

P6: shall we just do this on paper now?  
P2: a spreadsheet as it’s called...  
P3: a spreadsheet would be a good idea  
P1: If P6 can just put it up as a column and label it...  
[happens on screen]  
P1: You might want to put another column in, appearance, which we might not put anything in... [decide on who’s dealing with which cells]... potentiality has to be another column...
The spreadsheet is progressively populated on screen as pairs work and the spreadsheet is annotated with many different kinds of information, including website addresses, unknown qualities, Wikipedia pages, and so on.

What follows over the next three months following the meeting is a concerted effort to populate the CTO with work done in the main by individuals from the group. This process entails a number of Skype calls and one face-to-face meeting by a small number of the participants. Space precludes any coverage of a huge amount of individual and small group work done in this way. The main point, however, is that this development of the leaf nodes (the populating of the ontology) is, in the main, individual work. One of the interesting features of this is that the new CTO grows considerably. By the time the group reconvenes it consists of over 200 intermediate cell types (and 400 leaf nodes). This time, two more people have been invited, largely because they bring specific expertises to the group that were not previously present, and it is really now that the serious business of developing a ‘shareable’ subsumption hierarchy in the ontology gets underway:

P1: [we need to] check some of the biology and particularly our usage of the GO process ontology... we need to plan where we need to get to and in particular how we’re going to validate the normalised ontology artefact we’ve produced... so we developed a schema and set up a series of spreadsheets to describe the properties... and we filled out the values using various supporting ontologies like GO process, PATO, the cellular component ontology, FMA... What P7 has set up is a series of scripts which will take these spreadsheets and generate the OWL encodings and build the ontology by a pipeline... automatically...

The group begins by looking at contractile cells, information about which has been gathered by one of the group members. The work being done here is that of producing the hierarchy. Again, this work is complex, and involves both the resolution of ambiguities and decisions about the ‘best’ way to code matters in the light of evolved purposes:

P7: yeah, OK... this is it [on screen] start with the fast muscle cell... on the top you see annotations... I believe the process was put in by P4.

P4: yes, that’s one of mine...

P3: can I make very general comments... when we’re considering contractile cells... there will be certain cells which are clearly not muscle... hair cells in the inner ear used for hearing are known to [gestures] contract at high frequency... fibroblasts remodel the extra cellular matrix by contracting and pulling... so while a myoepithelial cell is a sort of muscle cell as well as sort of secretory cell there are others which are you can argue that are clearly not muscle that can contract so one thing we need to make clear you can be a contractile cell without being a muscle cell.

P4: I think that is... I think there aren’t many... but there’s at least one

Of course, this process also entails the identification and correction of mistakes. Sometimes these are easily agreed and rectified, but not always. Deciding upon what a ‘mistake’ is will not always be unproblematic. Firstly, there will be different kinds of mistake. For instance, some mistakes might be thrown up by the reasoner after decisions are made and agreed:

P2: could we just look at all the children of contractile cells?
P7: [runs reasoner to identify relationships].
P2: I just want to see all the child term leaf nodes of contractile...
P3: flight muscle cell, that’s interesting... no, a cardiac muscle cell is not a skeletal muscle cell!!
P8: a flight muscle cell is never a cardiac muscle cell
Regardless, corrective work is the main part of what is done at this late stage. As classification decisions evolve what was once ‘right’ may now not be; original assumptions may have been entirely wrong; there may be sins of omission, or poor or careless input work. In any event, corrective work is done by those who know:
P9: Pericyte... you’ve got it wrong... I’ve just been looking it up on the web... it’s been used here as an example of a single smooth muscle cell on a blood vessel... that is out of date, it’s now known to be a primitive cell form, undifferentiated... I found two references to this just now... it can differentiate into, one, a macrophage, a fibroblast or a single smooth muscle cell...
P1: So it develops into
P9: It develops into... I can give you the reference for this...
P1: how have we got it axiomatically described?
P7: yeah, it’s ‘located in’ blood vessels, ‘participates in’ angiogenesis, and ‘participates in’ blood vessel and ‘participates in’ organisation of an anatomical structure...
P1: so we’re saying all this is wrong...
The important feature of this, in our view, is that this corrective work is very much a product of the socio-technical distribution of expertise. Even the most expert of cell biologists may fail to recognise issues which are outside of their immediate area of interest:
P1: do we want to look at any other contractile cells?
P8: What about sea urchins?... I think one category is worth entering... and that’s epithelial reorganising cell... these are the cells that form the gut of sea urchins... it’s a subclass of epithelial... it’s just so interesting in terms of cell type development... you can take this little bit in culture and it goes whoompf before your very eyes...
P2: so, P8, I’ll put down that you’ll define it [laughter]
Biological work is frequently done through the study of ’model organisms’ – organisms that in some respect are either similar enough to human beings or archetypes for a type of organism to exemplify certain aspects of biology. Sea urchins are one such, frequently studied for purposes of understanding embryo development. What is being pointed out here is that the CTO does not include a class of cell that is important in such studies. For these purposes, then, the CTO would be viewed as ‘incomplete’.
In sum, the process of building an ontology in this instance, and we suspect generally, has been centrally, and iteratively, a matter of identifying the scope and ambition of the ontology. In the course of this, ad hoc decisions have been made about what work is best done collaboratively, how best to deploy the skills and expertises of members present and of others (and what those skills and expertises are) and how to obtain and use...
information. All of these matters, we think, have ramifications for distributed ontology building.

### 3.3 Visibility and Comparability - Artefacts and the development environment

The challenges of both comparability and visibility are nowhere more evident than in the artefacts that are used as ontologies are produced. This is evident with reference to the Protégé-OWL development environment.

Thus and for instance, when one colleague asked,

> So, our questions would be, does the default view in Protégé support the activities of users... some people are just browsing, some are building... a lot of users are just skipping over the fact that you can configure... it would be nice to provide more default views rather than plug-ins that people don’t use...

It immediately becomes obvious that the question carries direct implications for the development environment – a clear example of tools needing to be ‘sensitive to working conditions’. The point is that answering this question requires rather more than some scenic description of ‘community’. Indeed, this was brought home to us when we tried to gain a sense of the different things that any given user might want to use an ontology for and the prospect that they might be ‘just browsing’ was identified. ‘Browsing’ behaviour was of some interest to us and so we subsequently asked,

> So, do we know what people are doing when they browse?...

And received the following response:

> Not really... they might be evaluating this ontology against another one... they might be ‘delving down’ to assess the complexity of the ontology... they might be looking to see what domain knowledge looks like “on the borders”

The lack of knowledge about users might be critical. In another context (knowledge management) one developer explained their decision not to use OWL in the following way:

> P3 is such a big complex system. Ordinary computer users with a rapid prototyping background, they don’t understand OWL and the business of classes and individuals. In our case, we have domain experts working without knowledge engineers... doing the classification work [using folksonomic strategies prior to formal knowledge engineering work] as they do their day-to-day work. If you have a whole community working like that, well you have to work with the weakest. But the advantage is, this is the way people see the value... see the use. Protégé implies you’ve read the manual, you know what an ontology looks like. But we were looking for the sweet spot between tagging, the semantic web and formalisation. My own view is that different flavours of expressivity are needed in OWL... there needs to be a trade-off between modelling power and reasoning complexity which relates to communities of user so you can avoid this constant problem of developing and using at the same time.

There is a very considerable overhead to the learning of Protégé-OWL, one that many domain experts are not willing to undertake. When they do, it seems they cannot always easily get to grips with the various functionalities offered. Our study of the CTO-building group also revealed something about the enormous number of other artefacts that are
deployed during a process of this kind, and which have to do with the specific ontology being built on any particular occasion. We noted the use of a Wiki; several different existing ontologies; several versions of an Excel spreadsheet; a textbook; Wikipedia; pieces of paper; a flipchart; a whiteboard; SourceForge; OBOedit\(^{11}\) and a number of Google searches. It follows that, if these artefacts are all being used, it means that the work cannot be done more elegantly entirely within the Protégé-OWL environment. In itself, this is something of a challenge in terms of how these things are to be shared in a distributed work context, but, in reality, the problem is less to do with the number of artefacts than it is with the fact that they are often used simultaneously; often require specific skills, and are sometimes used in rapid succession. Frequently, there is an obvious need for a shared public view, while other databases, ontologies and information sources are independently searched and, at all times, there is a need for rationale recording on the Wiki, which had been prepared in advance. This was an entirely non-trivial part of the process, for it was obvious that participants needed a record of decisions made and their rationale in order to continue to do the work between the face-to-face sessions. The following extract from the CTO meeting gives a flavour of this:

P6: P1, I’ve just put the list up on the screen... I just extracted all the terms... there’s a thousand here... is it useful just to scroll down it?[on screen, P6. Navigates through]

P1: As we go through the screens, can someone have OBOedit open?
P3: yes, but how do you do search in OBOedit...
P6: you use term filter...
P7: so, we’ve got the list... [appears on screen with IDs]
P2: have you got obsolete terms in there as well...
P7: yes...
P2: better to invert them, cos the high numbers are likely to be leaf nodes...
P1: good point...
P5: course, now we’re going to have terms where we have no idea what they mean...
P2: Wikipedia man... trophectodermal cell
P6: No, there are no definitions for Trophectodermal cell... so not that one...
[they proceed down the list. P7 reads aloud]
P1: we can record these in the spreadsheet, P4.
P1: don’t forget the implicit categories...

Here, in the space of less than a minute, we see the use of a number of different artefacts. They are used synchronously, or in rapid succession, and more than once there has to be an exchange of information about how best to use them. We identify two features of this heterogeneous artefact use.

Firstly, it is evident that rationale recording is important. Now, we had previously seen this point made in a pedagogic context, when Computer Science Masters students were being taught to build an ontology:

\(^{11}\) OBO-Edit is an ontology editor optimized for the Open Biomedical Ontology (OBO) file format.
Organize them informally, paraphrase and clarify them to produce informal concept definitions... paraphrasing is really important... how will we know what you’re trying to do if you don’t make notes... photographs are good too... I mean it, if you have a digital camera or a mobile phone, take photos of the way you organise the cards... felt boards are useful tools for organizing things...

and,

without a paraphrase you can’t disagree on why we did something... what can we say about all members of a class?... all of this does some of that or all of this... these are the only two constraints we’ve got.

One of the things that became very clear to us in our observations of this course was that, although students were very adept at building ontologies from scratch, they were much less aware of the need for rationale recording, even after the instructions we outline. That is, although they were quick to understand the technical aspects of the work, they were much less adept at understanding the collaborative, interactional, features of knowledge work. The CTO group, who were much more experienced in the ‘real world’ business of ontology-building, evidently accepted the importance of this in that they maintained a wiki right through the exercise. The wiki was maintained by one person and this, we feel, was a function of the fact that the members of this group knew and trusted each other. The issue of how to record rationale in a more distributed context has not yet been resolved. This is not merely a problem of enabling annotation – it has to do with identifying what kinds of annotation are necessary and how they might be conveniently separated so as to be easily identified by the various categories of user. These issues tie in closely with the kinds of debate that inform the ontology-building community today – debates about problems of scale, scope, detail and usefulness.

Secondly, user practices and user expertises in respect of the tools used are vitally important. The very fact that a group of experienced, and expert, users of Protégé-OWL needs to use a variety of resources in the above example tells us something about the issues involved in distributed ontology building. It seems that artefacts need to be ‘at hand’ to recover specific information; that these artefacts are contained in a number of different electronic and physical forms, and that knowledge of how to use and get the best out of the tools is socially distributed (as in, ‘how do you do a search in OBOedit?’). Certainly, we do not know enough about why there is a preference for some tools (e.g. OBOedit) over others. It is apparent that many users are simply not aware of the availability of certain kinds of functionality within Protégé-OWL, for instance, and often have recourse to other, quite separate tools. It could be that the bigger the tool set provided, the more complex its use becomes for naïve users. Regardless, the heterogeneous nature of these artefacts-in-use has obvious implications for tool support in distributed collaborative contexts.

4. Conclusions
We have argued that ontology building can be understood as a complex of socio-technical issues. As such, treating them merely as ‘engineering’ or ‘modelling’ problems runs the risk of missing some important features of the process. These features, which have to do with interactional and ‘work’ processes in which mutual understandings have to be identified and specific knowledges interwoven, can be glossed as ‘social’ although
they closely relate to technical decision making. We have tried to show that ontology building - at least in the cases we have been able to study- seems to have some special features. Firstly, it raises the problem of classification work and how it might be understood. We argue that this, in turn, maps onto issues that can broadly be thought of as to do with Bowker and Star’s (op cit) notions of comparability, visibility and control but, in this context, translates in quite specific ways, notably to do with issues of enlistment, entitlement and purpose; defining method and scope, and the use of the development environment and other artefacts. Our work here has been focused upon elucidating the problem of ontology building in a collaborative context. After all, the issues of how to go about building, maintaining and reusing ontologies have been a long-standing concern, and it seems researchers in this area have long understood the need to focus on ‘how to do it’ hand in hand with ‘how it (really) gets done’ (see e.g. Gruber, 1995).

The first of these obviously has to do with features of the ‘user’. We have shown how an otherwise ‘technical’ problem - how to define terms and relate them in a classification hierarchy relates to assumptions about use and users. That is, in this context, the problem of comparability is at least in part a problem of determining the uses to which a putative community will put the ontology, and how the limits of those uses can be defined. These features, as demonstrated, have an immediate impact on the way in which an ontology is put together. Not least, it is evident that the notion of community is problematic in ontology building: in this context, ‘virtual community’ is probably more appropriate. The first problem is to arrive at a better understanding of the degree of heterogeneity in any given community, and how that might relate both to the kind of sophistication required of the ontology and the design of the tools with which it is to be done. This may include, for instance, a need for knowledge of the relative expertise and authority to be found in the community; of the closeness or otherwise of the relationship between ontology builder and domain expert, and of the different kinds of domain knowledge that may actually turn out to be relevant. This would not matter were it not for the evident fact that when groups of people get together in order to build ontologies, they spend significant amounts of time dealing with exactly this kind of issue and it seems, on the evidence we have garnered, that this has a direct impact on the size and scope of the ontology being built. Moreover, the discussions and disagreements entailed are precisely the aspects of the process that seem to necessitate ‘whole group’ activity.

Related to this is the fact that ‘users’ in ontology building are not related to ‘ontologists’ in any fixed way. In different groups, we see different kinds of relationship. In some, as with some of the bioinformatics work, ontology builders are typically biologists as well. Even there, the problem of ‘bona fides’ raises its head: who, that this, should be permitted to make authoritative judgements? This is no simple matter. In the context of ontology building, the familiar problem of the immediate day-to-day purposes of the user measured against the longer term, and meta-level, purposes of the ontologist are particularly vexatious, and magnified by the highly technical knowledges being deployed by all parties. Much of what we have seen demonstrates how the user remains a largely imagined vector of requirements, understandings and practices, with occasional contacts with and feedback from actual persons who might count as users.
The second specific challenge we identified was that of defining method and scope, which we characterise as an aspect of visibility and control. Software engineering has a range of approaches have evolved, from ‘structured’ methodologies to more ‘agile’ (see e.g. Beck et al., 2001) methods. In ontology building, the relationship between the two is less well understood. What is evident from our studies is that ‘methods’ are deployed, redeployed, adopted or rejected for entirely pragmatic reasons. A number of these methodologies for ontology building exist, of which the best known is probably ‘Methontology’ (Fernández-López et al., 1997; Fernández-López et al., 1999). Methodologies of this kind prescribe an order to a development life cycle and typically structure development according to a series of activities or tasks. On the basis of our studies, we would argue that they have little to tell us about the practical problems entailed in distributed ontology building.

As we have noted, problems arise even in a relatively homogeneous small group where all members are known to each other. Again, the striking feature of the work involved is that the effort that goes into defining method, scope and ambition necessitates a great deal of face-to-face involvement and that it is iterative. Our data suggests that there is substantially more work done on these matters than might have previously been recognised, that it is arguably the most vital part of the whole process, and that there are good reasons for thinking that this is done economically through a synchronous collaborative process. It is an activity that can be conceptualised as quite distinct from the business of populating an ontology (something that can seemingly be done without the same group work - in our data it was done by either individuals or in groups of two and three, and more often than not by e-mail.

The third challenge we identified had to do with a second aspect of the broad issues of control and visibility. This had to do with the artefacts in use during the process of ontology building, one of which is the development environment. Our observations of the CTO building process led us to identify a large number of different resources being deployed, as well as a fluid approach to the division of labour (individuals doing work, divisions into small groups, appeals to expertise and knowledge) and a rather ad-hoc approach to the structuring of activities (which includes identifying the ways in which changes in one ontology may impact on others). This eclecticism of procedures, technologies, techniques, etc., is manifested in the wide variety of resources deployed. As we have indicated above, however, it is not that they are deployed that is important, it is how. The problem for distributed collaborative work is evident when we look at the business of rationale-recording. We saw at many points in our observations, how important this is to the professional ontologist, again because it relates back to the issues of user, community and purpose discussed above. Our observations of the CTO group demonstrated that in a small group of trusted members this was left to one person to transcribe in a Wiki. Nevertheless, vast effort was expended in discussing and revisiting rationale. Equally, where some artefacts, like the Wiki, and like certain web resources, could be left to the individual, others could not. There is no doubt, as one ontologist put it, that ‘there are a lot of bad ontologies out there’ and much needs to be learned about the socio-technical distribution of expertise in each instance. ‘Expertise’ here – and this has been the point of our observations – is a complex problem since it refers both to evolving competence in the use of ontology tools and to knowledge about the various domains and their aspects that can be modelled (and equally importantly, cannot).
The process of ontology building is not well-understood, and distributed ontology building even less so. As one participant in our study put it, ‘the typical answer to the question, ‘how do you build a good ontology is, ‘the way we did it’. Understanding what issues arise during the routine course of ontology production, the order in which they are dealt with, and the way in which a combination of resources and expertises are deployed so as to evolve the desired ontology may ultimately help us provide support for that process. We believe that some key elements here have been largely unrecognised. Firstly, understanding what work is done face-to-face and in groups, and what work is more easily left to individuals or sub-groups to complete is important if we are to understand the work that makes an ontology ‘shareable’. We noted in this case that scoping the ontology involved agreements about the exclusion of whole cell-type categories because they were difficult, or because they were unlikely to be useful. We might note in passing that, at time of writing, existing support for collaborative ontology building (e.g. Collaborative Protege 3.4) does not have even rudimentary transaction management/ cursor control protocols to support such activities. Thirdly, a very significant part of the work involves the need to ‘test’ decisions. This involves the resolution of ambiguities and the correction of errors. We have tried to show that this is entirely non-trivial and is largely dealt with by the constant delaying of a ‘final’ version and through the collaborative deployment of a range of skills. Again, though we are not suggesting such work cannot be done at a distance, there is no question that here it is done elegantly and swiftly in a face-to-face context. How this might be achieved in much larger, distributed communities is not a settled question.

References


