What about Sea Urchins? Collaborative Ontology Building among Bio-Informaticians

What about Sea Urchins?: Collaborative Ontology Building among Bio-Informaticians

Dave Randall[1], Wes Sharrock[3], Rob Procter[4], Yuwei Lin[4], Meik Poschen[4], Christian Griefenhagen[3] and Robert Stevens[2]

[1]Department of Sociology, Manchester Metropolitan University, UK

[2]Department of Computer Science, Manchester University, UK

[3]Department of Sociology, Manchester University, UK

[4]National Centre for e-Social Science, Manchester University, UK

Email address of corresponding author: D.Randall@mmu.ac.uk

Abstract. This paper reports on an ongoing study of collaborative, distributed ontology building. Ontologies are now widely used in the biological sciences and are increasingly being deployed to other purposes. Here, our ethnographic study (see Randall et al., 2007) concentrates on a series of meetings and online discussions taking place with a small group of people meeting to rebuild a 'hand crafted' ontology. It aims to compare the results of a study treating ontology building as practical work, thus affording a comparison with more formal ontologies and offering some initial insights into the consequences for distributed ontology building.

Collaborative Ontology Building among Bio-Informaticians

Wikipedia defines an ontology as (*inter alia*) a shared definition of collections of entities from a domain, ranging from abstract concepts to specific instances and their attributes, with specified relations among the entities, and which supports automated processing and reasoning to derive new facts, properties and relations. It goes on to suggest that a typical ontology lifecycle consists of the initial creation of the first version of the ontology, its use and maintenance, its comparison with other ontologies and finally its, usually through integration in a wider ontology.

There are four elements of ontology building that we want to discuss in this paper. First, we aim to understand what work is entailed in the 'sharing' of definitions (see e.g., Bowker and Star, 1999; Ackermann et al., 2003). Second, we seek to identify how which expertises and authorities are mobilised. Third, we are interested in practical work insofar as it is 'left out' of standard methodologies for ontology building, such as 'Diligent', OTK and methontology (see http://www.aifb.uni-karlsruhe.de/WBS/cte/ontologyengineering/ for an overview) and how best to view that work (see Lin et al., 2007; Pike and Gahegan, 2007; Ernst et al., 2005). Fourth, and founded on an argument concerning the relationship between observations of real world work and design decisions, we will make some preliminary observations about tool support for distributed collaborative ontology building (see e.g., Buckingham Shum et al. (2006) on rationale recording).Our specific aim in this paper is to identify how the process of ontology development in this context unfolded in practice during the course of the building of a cell-type ontology (CTO).

The work we report on here was a *quasi-experimental* process whereby a group of bio-informaticians met (and continue to meet) over a period of time to reconstruct part of an existing cell type ontology (CTO). Such a usage may seem odd, but it describes the fact that, although the ontology building process was 'real world'– that is, it was intended that the product could be used in some way by biologists – it was also done with a conscious eye on the process itself. That is, it was an experiment only in the very loosest sense. Meetings were attended over a total of four days (so far) and video recorded by a sociologist who was also able to monitor e-mail communications and to interview individuals about their experiences at various times. Below, we describe the way the process unfolded.

Choosing an Appropriate Ontology

Initial decisions concerning what ontology might be used for this work were, on the face of it, very straightforward. The task had been conceived of by members of the Ontogenesis research group which consisted in the main of biologists. A group of people was invited to participate on the basis of expertise in cell biology and in ontology construction and, moreover, on the basis of 'people we can trust ... I know I can get on with them ...' Having said that, one feature of the work was the changing ambition and scope as the process developed over time. 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,

> 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. First, what exists is a 'tangle' which contains several different hierarchies. Second, what exists contains a number of errors and, third, 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 – derive 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.

Making sense of the 'community' – expertise and warrant

To begin with, some assessment of the current state of play has to be made, and in part this entails arriving at some judgement about what community of people is interested in this ontology and what commitments members of that community might already have made. As we see from this extract, this is a political process:

> R: ‘from what I understand is that the OBO[1] 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 ...’
> H: ‘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 ...’
> D: ‘who is ‘they’ here?’
> H: ‘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 ...
> D: ‘but that no man’s land is located over in Houston ...
> H: No not particularly, though most of those people are over in the US ... X is in Harvard ... right now, Y is in Boston 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 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:

> H: 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 ...’

Of course, there may be a number of purposes envisaged and, as we have pointed out, the ambition and scope of this work has not yet been defined. Indeed, we will argue that it only becomes defined in the process of ‘doing’ the work, and with constant reference to the interests in play.

How to proceed – doing ‘good’ work through disassembly and assembly

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:

> R.: ‘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

[1] 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

stage. Which supporting ontologies 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 very surprising aspects of the process, to the attendant sociologist, 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:

> *R: ‘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 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:

> L: ‘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 ...
> D: for use by biologists it would have to be the cell-type ontology.
> R: 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. 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. For brief mention, anatomists and medics might have interests in such an ontology. 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[2] 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 modeling 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, which models some cell types explicitly.’ The size, scope and ambition for this ontology, then, have to do with exactly who will use it. The laughter is occasioned precisely by the fact that different philosophies concerning ‘use’ and ‘truth’ are in play here.

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:

> L: ‘one thing, above that level, are we just looking at cells in vivo? Or experimentally modified?
> R: They just have like six crosses ... they say that this is very weak ... they just haven’t populated it ...

[2] CCO is the Cell Component Ontology

> H: 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 ...’
> L: 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 ...
> H: 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 ...
> L: I agree with R., 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:

> D: ‘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 ...
> R: ummm ...
> D: we don’t need to classify cell types in yeast ...
> R: we haven’t made that decision yet ... ummm ... we might have done [laughter]
> H: 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 which relate to what is ‘doable’ and to the possible value of the results. In the latter instance, for 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.

A principal axis of classification?

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 social distribution of expertise is critical when we look at the information getting strategies of the group later. A lengthy discussion ensues which involves the discussion of the various properties of cells which might be termed, ‘rigid’. Each in turn is raised as a possibility, some are dismissed quickly and others raised 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:

> H: ‘I think we may have got to the point where we cannot find a primitive axis ...
> R: 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 ...
> L: 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 ...
> D: we need types of function ...

> L: we need a process hierarchy
> R: which, funnily enough, we have in GO ... so are we happy that we just have cell and do it all by restriction?
> L: 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 that they wish to work with. 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:

> L: '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.
> H: 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 its 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 ...
> R: if we just go and pick 20 ... 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 20 ..
> H: we can do that collaboratively now ...
> R: 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 ...'
> [H. 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 Protege at this point, and a simple hierarchy with 'cell' at top and 25 cells listed underneath appears on screen in front of them:

> R: 'So, what ... we've now got 25 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, staininjg, 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 to that of another, but things are rather more complex than that. Firstly, and because ontologies have specific purposes, terms may be defined in different ways or may be incomplete:

> H: 'maybe you'd like to say what the problem is ... and I think it's 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 ...
> R: so its a watch this space problem ...
> L: the other thing we've learned is that PATO is seriously short of synonyms ...
> H: they're definitely missing synonyms they know that
> D: PATO hasn't been created with cells in mind ... its animal oriented ...
> H: its animal orientated because the people who are using it are using it to annotate mouse phenotypes ...
> L: 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 ...'

Second, it may turn out that the work requires some modification of an existing ontology, which will in turn necessitate liaison with other figures:

> (PATO is up on the screen) 'can we see it? ... it's got biological sex, then ... but here's nothing about chromosomal basis ...puts it up on screen ...
> H: so all of them come from the PATO ontology and those are in there by request of the project I used to work on ...

> D: male and female is all I would want ... let's get some procedure here, H. Will you undertake to request these changes to PATO ...
> H: 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 pato tracker which isn't something I want to do ..
> D: but you can cause PATO to evolve
> H: what I can do is go and see G. ... he's in Cambridge ... and then I'll report back.'

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'.

> R: '... the other thing is, if in the toy we've got 25 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 which contains a matrix of the selected cells and the properties that the group has decided to describe is distributed among members is developed and taken away:

> S: 'shall we just do this on paper now?
> H a spreadsheet as it's called ...
> D: a spreadsheet would be a good idea
> R: If S can just put it up as a column and label it ...
> [happens on screen]
> 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, etc, etc, Wikipedia pages, and so on.

What follows over the next three months 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 at least one face-to-face meeting. 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:

> R: '[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 M. 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:

> M: ‘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 H.
> H: yes, that’s one of mine ...
> D: 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.
> H: 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 they 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:

> H: ‘could we just look at all the children of contractile cells?
> M. [runs reasoner].
> H: I just want to see all the child term leaf nodes of contractile ...
> D: flight muscle cell, thats interesting ... no, a cardiac muscle cell is not a skeletal muscle cell!!
> E; 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:

> A: ‘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 ...
> A: So it develops into
> A: It develops into ... I can give you the reference for this ...
> R: how have we got it axiomatically described?
> M: yeah, its ‘located in’ blood vessels, ‘participates in’ angiogenesis, and ‘participates in’ blood vessel [] and ‘participates in’ organisation of an anatomical structure’ ...
> R: 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 social distribution of expertise. Even the most expert of cell biologists may fail to recognise issues which are outside of their immediate area of interest:

> R: ‘do we want to look at any other contractile cells?
> E: What about sea urchins? ... I think one category is worth entering ... and thats epithelial reorganising cell ... these are the cells that form the gut of sea urchins ... its 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 ...
> H: so, E., I’ll put down that you’ll define it.’ [laughter]

In sum, the process of building an ontology in this instance, and we suspect generally, has been one of identifying the scope and ambition of the ontology in and through the work of building 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.

Conclusion: Artefacts in Use

The data described above shows, to some degree, the wealth of artefacts that get deployed during the course of this work. 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 and a number of Google searches. In itself, this is something for a challenge in terms of how these things are to be shared in a distributed environment, 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; are sometimes used in rapid succession. Frequently, there is an obvious need for a shared public view, whilst 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 gives a flavour of this:

> S: 'R, 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, S. Navigates through]
> R: As we go through the screens, can someone have OBOedit open?
> D: yes, but how do you do search in OBOedit ...
> S: you use term filter ...
> M: so, we've got the list ... [appears on screen with IDs]
> H: have you got obsolete terms in there as well ...
> M: yes ...
> H: better to invert them, cos the high numbers are likely to be leaf nodes ...
> R: good point ...
> L: course, now we're going to have terms where we have no idea what they mean ...
> H: Wikipedia man ...trophectodermal cell
> S: No, there are no definitions for Trophectodermal cell
> ... so not that one ...
> [they proceed down the list. M reads aloud]
> R: we can record these in the spreadsheet , H.
> R: 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.

The process of ontology building is not well-understood, and distributed ontology building even less so. As one respondent put it, 'the typical answer to the question, 'how do you build a good ontology is, 'the way we did it'. Understanding what problems arise during the course of ontology production, the order in which problems are dealt with, and the way in which a combination of resources and expertises are deployed so as to evolve the desired artefact may ultimately help us provide support for that process. We believe that some key elements here have been largely unrecognised. First, 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'. Second, our data suggests that there is substantially more work done on defining scope and ambition than might have previously been recognised, and that there are good reasons for thinking that this is done economically through a synchronous collaborative process. 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. Third, 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.

Acknowledgements

We would like to thanks the National Centre for e-Social Science for supporting this research. The work in this paper could not have been done without the active cooperation and support of the Ontogenesis Network, funded by EPSRC Grant number: EP/E021352/1

Bibliography

Ackermann, M., Pipek, V. and Wulf, V. (eds.) (2003). Sharing Expertise: Beyond Knowledge Management, Cambridge, Mass., MIT Press.

Au, A., Xiang, L. and Gennari, J.H. (2006). Differences Among Cell-structure Ontologies: FMA, GO, & CCO. In Proceedings, Annual Fall Symposium of the American Medical Informatics Association, pages pp. 16-20.

Bowker, J. and Star, S. (1999). Sorting things Out: Classification and Its Consequences. Cambridge MA: MIT Press.

Buckingham Shum, S.J., Selvin, A.M., Sierhuis, M., Conklin, J., Haley, C.B. and Nuseibeh, B. (2006). Hypermedia Support for Argumentation-Based Rationale: 15 Years on from gIBIS and QOC. In Dutoit, A.H., McCall, R., Mistrik, I. and Paech, B. (eds.) Rationale Management in Software Engineering. Berlin: Springer.

Ernst, N. A., Storey, M.-A. and Allen, P. (2005). Cognitive support for ontology modeling, International Journal of Human-Computer Studies 62(5), 553.

http://www.aifb.uni-karlsruhe.de/WBS/cte/ontologyengineering/

Lin, Y., Procter, R., Randall, D., Rooksby, J. and Sharrock, W. (2007). Ontology Building as Practical Work 3rd International e-Social Science Conference, Ann Arbor, Michigan, USA.

Pike, W. and Gahegan, M. (2007). Beyond Ontologies: Toward situated representations of scientific knowledge, International Journal of Human-Computer Studies.

Randall, D., Harper, R. and Rouncefield, M., (2007). Fieldwork for Design. Berlin and London, Springer.