Engineering the scientific corpus: routine semantic work in (re)constructing a biological ontology

Engineering the Scientific Corpus: Routine Semantic Work in (Re)constructing a Biological Ontology

Wes Sharrock
University of Manchester
(wes.sharrock@manchester.ac.uk)

Dave Randall
Manchester Metropolitan University
(d.randall@mmu.ac.uk)

Christian Greiffenhagen
University of Manchester
(cg@manchester.ac.uk)

Introduction

In face of the burgeoning interest in 'ontology' in science studies, Michael Lynch (2008) called for a move toward 'ontography', to talking about ontologies by way of studies in which ontologies (or at least, an ontology) are of demonstrable relevance to the doings of those being studied.

This paper provides an ontography, or some part of one, in that it reports on work in ontology development being done by a group of researchers in bioinformatics, drawing its examples largely from a workshop in which some members of that group were participant and which was organised by a research network to which they belonged. Methodologies for building 'good' ontologies were part of the interests of this wider research group and were a motivation for the work undertaken. What is evident from our study is the fact that methods to be applied, avenues to be explored and even fundamental purposes were all in the event 'up for grabs' and formed a closely interlinked and mutually explicating part of the 'logic in practice' deployed.

We will show how this research work was undertaken with reference to an existing body of knowledge, yet requiring distinctive courses of 'discovering work', concerning both method and substantive content. How were its results examined and re-examined in the light of ongoing, evolving and unanticipated considerations? Describing how the involved participants go about their work is, then, 'an ontography' in precisely the sense that Lynch proposes.

Background

Social studies of science having been focused on how new contributions gain acceptance with scientific communities have tended to take for granted the fact that natural science investigations are conducted against a background of accepted and settled findings, without giving much emphasis to the way that current work is embedded in the accumulated results of prior work (for an exception, see Sormani, this issue).

With the proliferation of biological research and the contribution that it makes to the online 'data deluge', concerns about how to organise, manage and access existing data are increasingly becoming as significant as those of adding to the stock of biology's findings, though the two can be connected insofar as it is anticipated that ready and effective access to biological data stored on line can provide a potent opportunity to develop biological ideas and generate new findings without having to create new data. Currently, such ideas are largely aspirational. There are

developments in the organisation of the World Wide Web and other technologies that make Merton's Holy Grail of 'scientific communism' look more realistic than ever, but as yet these facilitating technologies have not been used to develop the computational infrastructures that will fully realise this.

'Finding out' in numerous forms, including the making of discoveries, is the front-line work of the natural sciences, and, as is well known, but relatively seldom documented, front-line achievements are commonly extensively and in detail, dependent upon support activities, some dedicated to a specific project, some given over to developing infrastructures that can support one or more genres of investigative scientific practice. Like other forms of work, scientific work is saturated with assessments of productivity, ranging from scientists' assessment of the productivity of scientific exemplars to administrators' concerns with the returns – however defined – yielded by forms of scientific effort.

Relatively recent and continuing developments incorporating computing into scientific work has spurred thoughts about the economy of scientific investigation, particularly in respect of the ever expanding number and diversity of scientific investigations, and as to whether the tying of scientific investigations to data-collecting efforts is the most productive form of inquiry in times when, increasingly, it is possible to enable the sharing of data, which in turn facilitates the re-use of existing data, which re-use may service discovering work through the methods of computer simulation.

These conceptions are to some extent already a reality, but there are expectations that their full development could quite profoundly transform the nature of investigation with computational operations assuming greater importance than theory or experiment as a means of discovery. This paper attends to some work devoted to infrastructure development through the formation of on-line technologies and tools that could, long term, facilitate the re-use of data for discovering work (mostly – since the work is bioinformatics – in respect of biology). This work is understood as a part of a much longer term development, as relatively early work in the development of ontology building capacities that will eventually enable use of the web for the storage, precise identification and recovery of all data relative to some specific scientific problem.

The bio-informatic work is itself innovative in terms of developing general methods of ontology building to high engineering standards. One main aspect of their work is knowledge representation through capturing a logical structure for some domain of classification. Another aspect of that work is finding out about the range of understandings associated with categories used in the relevant domain so as to identify some recognisable consensus of usage that can be incorporated into the representation. The workshop used to exemplify some of the bioinformaticians' work was explicitly nominated by its convener as a means of sociological discovery, which was one reason for its being recorded. The workshop, involving an exercise in collaborative ontology building, was a proxy for online ontology building - it did not itself involve online collaboration - but it was a means of discovering something

about how collaborative ontology building works, what practices it involves, what supports it needs, what resources it calls upon and so on.

Our case: a bioinformatics workshop, ontology-in-practice and science studies

The Network Workshop involves bioinformaticians and biologists (who are often the same persons) engaged in such infrastructure work by specialising in the development of 'ontologies' which they regard as an essential step toward providing developed semantic content for Web 2.0[1] relevant to the identification and recovery of biological data stored on line.

'Web 2.0' has been associated with ambitions to change scientific and engineering practice. After all, in principle, a world where collaboration is possible at the click of a mouse means that scientific work can be done in a quicker and more efficient manner. At exactly the same time, and in another feature of Web 2.0, the 'semanticization' of the Web means - again, in principle - that the prospect of a standardization of concepts is possible. If that proves to be the case, this ought to have benefits for the practitioners involved. It means, inter alia, that data from whatever source can be used and re-used for comparative and other purposes, further enabling 'scientific communism' in and of sharing data.

Though there are some relations between 'ontologies' of the kind being built by computer scientists and the traditional philosophical ambitions for a scheme that comprehensively identifies all the kind of things that there are, these are complicated and organisationally remote from the kind of transactions we will be examining. Ontologies in the relevant sense are mainly concerned with sorting out the organisation of collections of terms current in some knowledge domain, and primarily for purposes of enhancing the computer processing of knowledge expressed by those terms. More specifically, an 'ontology' is an organised terminology made up of terms drawn from one 'domain' or another, structuring (some of) the terms from the language in use in that domain (terms for biological cells, or terms for the items stocked in a museum, etc., ad infinitum) to provide an arrangement that is sufficiently well organised that it can be used in search engines to make consistent and accurate identifications of online holdings of - in many cases - data, in this workshop's case, of biological data.

Much of the structure that is being assigned to the biological categories is being regulated by the 'first order logic' (a formal logico-mathematical system) that is programmed into the technology that participants are using as their main ontology building tools. The use of first order logic in these tools means that each class of objects has to be precisely defined and its relationship to other classes of objects in the same domain equally well-defined with respect to entailments so as to permit the automation of inferences.

At its simplest and crudest, the studied bioinformaticians' overall efforts can

[1] 'Web 2.0' is the catchphrase for developments in Web design that substantially increases the opportunities for 'user generated content' and the formation of online collaborative activities (social networks, video sharing sites and the like instantiate these developments).

be understood as attempting to manage two sides of their operation. On the one side, there is that of designing an ontology with a worked out, through-and-through logical structure, but ensuring, at the same time, that the terminology being incorporated into the structure has an adequate relationship to the vocabulary in use amongst biological professionals, where the meanings of terms is not necessarily uniform or unchanging. That is, the members of this group are doing work that involves both the deployment of logic and taking decisions about the definitions of terms – that is, semantics.

As numerous studies of the history/sociology of science and technology have shown, in principle does not mean in practice. There may be any number of obstructions to this vision of cooperative working. Indeed, science and technology studies (STS) have demonstrated many times how features of organizational and/or political infrastructures may prevent the easy realisation of these possibilities (e.g., Lee, Dourish and Mark, 2006; Bietz, Baumer and Lee, 2010; Ribes and Finholt, 2007; Ribes and Lee, 2010). Some of these studies show that divisions in science sometimes have to do with the development and installation of classificatory schemes (Bowker and Starr, 1999; Randall, 2001). Less often comes the recognition that scientific workers of whatever kind may both be oriented to these political and organizational limitations and may actively seek to overcome them. In the case considered here the participants understand themselves to be working on the formation of infrastructures for discovering work in the sciences more generally, though one focus for some of their concerns is the possibility of using ontologies to manage the data deluge in biology. They recognise that problems of classificatory diversity and even of conflict are ones they need to handle if an ontology is going to have sufficient utility to be taken up within the research discipline.

Below, the authors look at a series of specific examples of practical orientation to classification problems from 'ontology building'. Ontologies are seen as one possible remedy to classification problems. The data set is taken from two three-day meetings which bracketed some months of email and telephone correspondence.

We do not want to suggest that 'ontology building' in which the research group is engaged is a matter of taking a developed technology and applying it to produce specific ontologies on demand (though these researchers have a sense of themselves as, in some of their work, trying to service the interests of specialised groups such as biologists or medical practitioners by assisting them to build an ontology). Rather, they are engaged in attempting to develop the technology itself, to apply engineering methods to improving practices of ontology development and design. The actual construction of ontologies is, from their point of view, commonly a troubled work, often undertaken by those with little experience in and limited understanding of how to put an ontology together. As a result, there is extensive variation in the quality of the existing ontologies that can already, and in large numbers, be found online, ranging from those that are little used and of little use, to ones which are relatively well-engineered.

Among the latter the best known and most used are ontologies put to specifically biological uses. They include, for instance, the Gene

Ontology (GO), the Phenotype Ontology (PATO) and the Chemical Entities of Biological Interest (CHEBI). All of these are part of the Open Biological Ontology (OBO).

Members of the research network understandably take a positive view of ontologies, but they are aware that there is controversy over the value of these, and over the philosophical foundations of ontological structures, not least over the extent and nature of relations between philosophers' and engineers' conceptions of ontologies. One of the problems they understand to afflict ontology development quite generally is the difficulty of building on existing work rather than having to start all over again. The methods applied in the workshop are intended as an address to this problem. The workshop is re-engineering existing ontologies which are recognised as ones that are important within biology but with logical structures that are open to significant improvement.

The practice: ontology building in and as ontology articulation

The engagement with ontology building amongst those we studied was a result of the development and combination of, importantly, three kinds of computational resources: the "Web Ontology Language" (OWL), semantic editors such as the one – "Protégé" – some of the workshop group were involved in developing, and semantic reasoners, forms of software designed to generate logical consequences from inputs. They were using these tools in combination to create a more methodic basis for ontology construction, now at the stage where a claimed virtue of their procedures was that ontology builders need not constantly 'start from scratch' but could build on and adapt existing ontologies or integrate parts from one ontology into another.

Much of the work being done on ontologies by the research group is strategic to the development of methods for ontology building, since it thinks of itself as possessing tools and techniques which can make the sound assembly of ontologies a more disciplined and dependable matter (thus more effectively facilitating its "heuristic" upshot). The group is heavily involved in methods development. It is not, of course, undertaking the development of ontology building methods from scratch, for ontology building has been a widespread activity in the computational sciences for some time (their disciplinary environment of biological sciences is already populated with numerous ontologies). Instead, as suggested, it is adopting and adapting tools which make both the formation and systematisation of fully worked out methods a possibility. Their work projects are strategic in the sense that tasks are developed on the basis of identifying areas of ontology building where there are no methods for forming ontological structures or where there are only cumbersome or partially worked out ones.

They make attempts to form novel methods for effectively and thoroughly dealing with problems in ontology building that are recognised as routinely producing sub-standard ontologies. Their work is very much 'tool centred' in that many of the tasks that need methodising are identified by contrasting what has been or can be done with other tools with what can potentially be done with theirs, and so their investigations are not so much 'how can this be done?' as 'how can this be done in OWL-Protégé'. These bioinformaticians are also advocates

for their tools, both through demonstrations that these can offer effective solutions to more general problems in ontology building and through providing open instruction as a form of capacity building for the use of the tools, where the value of the tool is understood comparatively – and not necessarily invidiously – in relation to other ontology building aids.

Many of their tasks provide projects for specifically dedicated research teams, but the instance that provides our example here is the product of an academic research network. As we've discussed at length elsewhere (see Randall et al., 2011) and will briefly need to consider below, their tool and method development naturally involves an *orientation to users*, possible and actual, with the researchers, often being biologists as well as informaticians, to some extent acting as users themselves, which is what they are going to do in the network meeting discussed shortly.

Nevertheless, their deliberations entail working out *in situ* what users, for what purposes, need to be considered in order that the ontology can be redefined and adequately circumscribed, for ontologies need boundaries in order to be 'usable'. In so doing, this group will attempt to specify the limits of the ontology by, inter alia, specifying what use it might be put to and working out what relation it might have to other ontologies in the OBO Foundry. In essence, the OBO (the Open Biological Ontology) provides a good example for them to work with, since it involves biology. Also, it features sub-ontologies such as the Cell Type which is of a manageable size for the task they have in mind, and the less-well developed aspects of that ontology provide an opportunity to try out their methods.

Examples: methodising and normalising CTO

We illustrate the paper by looking firstly at the beginnings of a workshop (projected as the first of two) to conduct 'an experiment' in 'collaborative normalization' of the Cell Type Ontology (CTO). The experimental status assigned to it reflects the fact that it is as much an exercise to find out what needs to be done in normalising an ontology as it is to improve the Cell Type Ontology as such. Subsequently we consider some of the normalising work done in stabilising findings, ironing out inconsistencies, reworking policies on inclusion and exclusion, and so on. The parties are all 'biologists of some sort' (except Dave Randall who was present more as observer than participant) but most of them are involved as bioinformaticians and are familiar with the general ontology building tools that are to implement 'the process of normalisation'.

The selection of the CTO is not due to any specific significance or priority of cell typing in biology, but because the ontology is a good candidate for this kind of exercise - it is, according to the convenor of the group (P1), 'hand crafted' (by a small number of people) as opposed to systematically constructed by some process of community consensus and it is to be expected that there will be 'errors' (which most likely intends 'inconsistencies') that are a product of its method of construction (hand crafting), which expectation is documented by P1's prior work[2]:

[2] These transcripts are edited. We have sometimes used square brackets ([ ]) to indicate more substantial excisions.

P1: *I want to take the OBO cekk type ontology which is a hand-crafted er taxonomy of cell types [ ] it's a multiple hierarchy its er what AR would describe er as a tangle*

P2: *( )horrifying*

P6: *It isn'tr pretty*

P1: *what tends to happen when you build ontologies by hand is that you will will 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 er modules reusable modules more maintainable lumps of hierarchy a- and more highly axiomatised ontologies what are*

P2: *wha*

P1: *highly axiomatised er*

P2: *yes yes*

P1: *there's more stuff in them more stuff in them so that you can make more computational inferences and essentially it does all the work.*

A major finding was that there are substantial failings, i.e. in one tenth of cases, where the 'subsumption relation' (which is a key relationship between categories and subcategories in the hierarchy of categories that is the backbone of the ontology) is either unspecified or wrongly stated. The CTO is also selected, as P1 explains in response to P2's direct question 'why the CTO?' because it has features which make it manageable for the projected exercise - it is small enough, its domain is focussed, there is an expert in cell biology in the group (and more join at a later stage), etc.

Methodising ontology building: Logical engineering

As explained above, the problematic for this group of researchers is not the construction of ontologies as such, but of *eventually* enabling the routine construction of 'quality' ontologies. Their work is not centrally focussed on meeting the demand for ontologies from any particular constituency, but concerned with developing, standardising and automating the skills of ontology building.

As the convenor indicates early in the meeting, "the answer to the question, how do you build good ontologies is usually 'the way we did it'". They seek, in other words, to contribute to the testing and assessment of some of their relatively untried development methods – that of 'normalisation' being prominent in this case. In the first instance they begin by deploying someone else's software, the "OntoClean" methodology (Guarino and Welty, 2002). OntoClean embeds a series of logical principles which can operate to force a more logical structure into the design of an ontology's structure, not least by providing means of perspicuously surveying the structure of the ontology-so-far and ordering it by defined principles. Such a method should allow for the definition of clear 'primitives' in the ontology and for clear subsumption hierarchies to be effectively derived from them.[3]

[3] Primitives are the 'roots' of any given ontology and cannot be defined by relating them to any other part of the ontology, except through axiomatic relationships with other primitives. Primitives in one ontology are not necessarily primitives in another.

What has happened is that P1 and a colleague (P4) have previously run the existing CTO through a computerised reasoner in order to highlight the omissions and logical inconsistencies that are being spoken about. Reasoners are extremely important in this kind of work because they are a means to check work done, and it is one maxim of the work that checks for consistency should be run frequently when assembling an ontology to pick up on inconsistencies as they are introduced into the ontology's structure and before further structures are built on them. This maxim reflects two other well-known features of the work: first, that those making entries without automated support won't be able, unaided, to track the consistency of the entries they are making and thus will unavoidably introduce anomalies into the hierarchy they are designing, and, second, that it is much easier to repair design errors before other structures have been built on them.

Though the members of the group have some understanding of ontology building, this knowledge is unevenly distributed and some members of the group raise a number of questions that may help them understand what they are doing and how to do it. Thus:

P1: *from what I understand and P6 might well know more is that the erm OBO people have commissioned a reworking of the cell type ontologyerm and I am perfectly happy for this to be a contribution to that process but that is not something I will manage... cause the whole process embodied in that would just drive me up the wall...*

P6: *so maybe I could comment very briefly on that we've been using the cell type ontology [ ] before this workshop 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 back to the root that they're addressing [ ] They had a discussion about should they rebuild the whole thing again from scratch take out all the hierarchy and just start again.*

P2: *who is 'they' in this context?*

P6: *people active are [gives a list of names]. The CTO doesn't have like a paid person to look after it so originally it was [gives other names] and now it's just sitting there in no man's land...*

P2: *but that no man's land is located over in [location]...*

P6: *No not particularly, though most of those people are over in the US X is in [location]... right now, Y is in [location] too... I don't know where Z is...*

There is here a recognition that work may well be ongoing on the part of the 'owners' of the CTO, work which will have some consequences for what this group is undertaking, especially as P2 indicates that the work may have 'real world' and immediate benefits. It becomes evident that understanding the nature of the existing CTO development and user communities' commitments has implications for the group's own purposes. Initial work, then, involves an assessment of the 'state of play' with CTO. That implicates a set of practical interests in addition to the implied political considerations of the above:

P6: *I would like to use the cell type ontology for my own uses n one one*

of which is tying all of the available public cell lines that we have data on and getting a type for them in the cell type ontology

P2: *Yes yes*
[
P6: *and making something cross-product which is really something that needs to be done and if we do that that could be something that would be really useful and that's something that [gives name] and I have sort of somewhere on our list of what we need...*

It progressively becomes clear that issues of method cannot be dealt with independently of substantive issues concerning the scope, size and boundaries of the ontology, all of which relate in turn to purported usage and which have to be practically fixed relative to the work in hand.

There are strong sociological tendencies to be suspicious of talk of logic (because of its association with doctrines about rationality, among other reasons) and, in the extreme, to treat it as an extrinsic feature of reasoning that is invoked only in retrospective justification of courses of action that were actually assembled without concern for intrinsic logical structure. Such arguments are a legacy of the idea that formal logical schemes represent the general process of thought, and a reaction against it, with the result that questions about the role of logic are treated as if they are *a priori* ones.

This paper does not propose any general view of the nature of logic but follows, rather, the precedent of Harold Garfinkel who repeatedly recommends to treat terms such as 'logic' to a large extent as place-holders for an array of as-yet-underdescribed activities, and this paper is an exercise in looking to see how 'logical reasoning' is done in an actual case, with the participants in the workshop facing two interdependent questions, how to set up a logically tidy general structure for their ontology as a whole, and how to site instances effectively within that structure.

Normalising built ontologies

An ontology, in its simplest terms, can be understood as an ordered construction of categories that are intended to capture a domain of phenomena and, at the same time, or, much the same thing, to express a form of expertise (compare with definition on pp. 106-107). This work is, after all, understood as a branch of 'knowledge engineering' or, as we treat it here, logic-in-practice.

This explicit orientation to problems of logic-in-practice is evident in discussions about method and, in particular, the stabilisation or, as we shall see, the "normalisation" of its results:

P1: *What I'm hoping that we will be able to do in identifying the primary axis is actually do this somewhat fo- formally and use one aspect of something called Ontoclean [ ] So what Ontoclean does [ ] but whats- erm ontoclean is a way of evaluating erm particularly subsumption relationships in ontologies and checking that you've said the right thing in the right way and it talks about unity, rigidity and identity and er unity is all about whether you're talking about parts and wholes cos one of the common mistakes is to talk about erm part-whole relationships as 'is a'*

relationships er erm 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 ummm ah ah th- th- the essence of things so wha er what are properties held by an entity for the duration of its existence or only for part of its existence. So I'm a person from the beginning of my existence to the end of my existence but I'm a student for only a portion of that time [ ] And what we want to do or what we should do is identify the primary axis of classification to be a rigid property and helps us we are told to make a nice safe tree

In these comments, formal logic is very much in use as a tool providing the initial procedural basis for the work. The objective of their methodological efforts is to provide a means for assuring a through-and-through logical structure for built ontologies, and hence to provide means that facilitate the further through-and-through logical expansion of the specimen ontology that they are proposing to rebuild. The problem is the first one mentioned by P1, that of 'tangle', and that is to be addressed through a process of 'normalisation'.

Local disentangling

One of the forms of logical structure that is central to the categorical hierarchy is the notion of 'inheritance', whether the properties characteristic of units at one level of the hierarchy are also possessed (inherited by) lower levels in the hierarchy. It is easy to see that many things can be simplified, for many purposes, if it can be assumed that many things remain fixed (and that no new things are added) throughout movement up and down the hierarchy. It is a feature of the language, however, that a subcategory has relations to more than one superordinate category, and it is a perfectly reasonable thing, if attempting to organise the collection of terms in a domain - this group often use Pizzas as a training example - to attempt to express all the relationships between categories in a single hierarchy, and thus to associate a subcategory with more than one superordinate category.

One of the standard engineering methods that the team are using is that of 'modular' construction, the design of parts of the overall construction so that they are extensively independent of one another. Modularisation is understood as a way of facilitating maintenance of the developed system by enabling changes to be made that can be contained within the module without ramifying throughout the whole structure. Here it is the lines of inheritance that are being treated as desirably modular units, with subcategories being associated with only one superordinate category so that the subcategory's relation to other superordinate categories can be assigned to segregated modules.

The failure to make such separations and the work of decomposing a hierarchy that has not been systematically modularised (that of CTO, in this case) results for the need for what they call 'normalising'. The point of such normalising (as indicated above) is that, when properly done, it enables reasoners (or people) to present clearly defined and unilinear subsumption hierarchies. It is, put simply, 'untangling' work, decomposing the CTO's hierarchy into independent lines. So the workshop group is trying out and learning about

modularising in Protégé OWL which will also be facilitated by learning about how their proposed methodology and the use of their tools can service the task in hand.

Attention to the structure as a whole is also manifested in considerations of how to start on reworking the existing structure. An initial concern is to find an axis which will function as the 'trunk' of the tree-like structure. In effect, this means trying to find some property which all cells have in common such that they can be defined as cells in the first place, and subsequently organised into subordinate 'types'. This is undertaken 'conversationally' in the sense that participants have to think about whether they can identify such a characteristic. This leads them into considering what kinds of 'cells' they want to include within their domain – do they want to include cells created in the laboratory rather than only those occurring naturally? Do they wish to include primitive cell types associated with yeast? They decide against this on grounds of time and competence, as shown below.

Negotiating classification

In some ways, the ontology that they are reworking can be considered a completed construction – it is an ontology of cells, but, as can be seen, there is room for negotiations about how complete the CTO needs to be. Decisions, that is, need to be made about 'where to start'. First decisions relate to what to include and what to exclude, at least for the workshop's purposes: what sorts of cells need to be included, what sorts of cells can practically be included in relation to the workload. Placement of the tasks is in relation to both the embedding logical structure and the organisation and schedule of the building team. Thus:

> 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 ...*

So, there are early examples of decisions about what to include/exclude from the ontology and why, and about where to place things within the ontology. A series of ways in which organisms can be generally classified by looking at high-level ways of characterising the properties of cells is proposed:

> *P2: cell by organisim could be a candidate for one axis*
>
> *P3: cell by function ...*
>
> *P2: Cell by histology, a classification by their microscopic- We think this is incomplete and what*

cell by histology means it's basically morphology or stainability

The point about this is that the identification of axes is a way of rectifying the 'tangle' of logical inconsistencies uncovered by the reasoner. The original version of the CTO, it seems, does not spell these, and other, properties out. It is complained, for instance, that in the prior version, cells are listed as 'mature' and 'immature' which does not in and of itself provide for any properties which might otherwise distinguish them.

Restrictive inclusion

The consequence of the above ambition, using the metaphor of the tree structure, is that the 'leaves' (i.e the identifiable kinds of cells finally subsumed into the classificatory system) would be the least difficult part of the logical work. If the main structure of axes can be derived then populating it should be easy (or at least, can be the result of empirical work, not logical work). Suffice it to say that the search is unsuccessful. No rigid property can be found. This means that a different method needs to be applied:

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 cell type ontology 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, funnily 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...*

The group looks for an alternative way of deriving a viable structure, envisaging different procedures:

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 it's all there... and then there's organism...*

[P2. Notes them all down]

At the lower end of the ontology there are 'the leaves' (i.e. the cells themselves) and the decision now is to take 'twenty or so' representative cells in order to populate the ontology but without trying to establish any significant degree of hierarchy (which means that little or no automatic reasoning can be done). Doing the work 'by restriction' entails a differently ordered kind of logical work. Here, what will be attempted is the classification of cells by defining certain kinds of logical relationship they have[4].

The point about representativeness is important, in that 'writing restrictions' is a way of identifying the individual members of a class in terms of the properties that they have, not by enumerating the individuals, but by specifying a property that they need to be counted in the class. Doing this will also require attention to what already exists. The group is aware that some of these relationships are already defined in other ontologies and they will need to interrogate them:

P1: all connected are we? 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, staining, lineage, anatomy, but we'll put anatomy to one side.

P2: you wanted the list of cross products

P1: yes, supporting ontologies

P2: morphology, process, nuclear number, ploidy, lineage, organism, size, maturity, anatomy, sex, embryo, proximity, location, potentiality...

P1: can we sort that list into PATOs?

P2: doing it now

P1: 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 for instance, it used to be the case that ploidy was just a flat list and maturity and immaturity might just not be there So when we go through the cell types, we might look at components but for the moment just go through the processes look for things like insulin secretion in GO.

What to do with cases?

Though the participants are biologists and have varying degrees of familiarity with the tools in use in the workshop, there are numerous occasions for

[4] For a more complete description, see http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_1.pdf

discussing and deciding about how to enter cases into the accumulating collection of categories. An example of bears on the nature of the relation that the scheme will say that two kinds of cells have to each other:

> P3: *here's a lot of thought into development A → B and C. Can you comment? Stem cell divides to be a stem cell and a daughter cell that differentiates - myoblasts fuse to be a multinucleate muscle fibre, process of change has been considered.*
>
> P1: *if I am correct all blood cells from hm stem cells may want to say hm stem cell → erthyrocyte, assumes that all → at least one erthrocyte, not true want to say it the other way around, erythrocyte develops from hm stem cell. We need a discussion on whether stem cells are immortal in this context.*

The issue here is about how to express relations of succession where one thing changes into two (A ARROW B *and* C) which query is exampled by a stem cell that divides into two, another stem cell of the same kind and a different kind of cell to that one. P1, the primary specialist in ontologies present, does not have a direct answer to the question but presents an issue which is, effectively, that of how the criterion of 'same' is to be used in such a context. Is the stem cell which results from the division a different cell from the initial stem cell: hence 'whether stem cells are immortal in this context', i.e. whether one of the two daughters of the stem cell, which is itself a stem cell, is to be counted as more of that initial cell or something different from it? This is not a matter to be decided at independently of other decisions about the structure of the hierarchy, in this instance on relating to hierarchically superior dimensions, as the following exchange suggests:

> P2: *is this a question of temporal processes and how we model those?*
>
> P3: *no more about modelling change, RS said that he is a person, and was a student, how do we model a myoblast that has become a multinucleic muscle fibre*

Commonplace examples such as that of 'is a person and was a student' are regularly appealed to in explaining the idea behind classification arrangements, so that there are general issues about how to treat cases where one thing changes into another, with the example of someone becoming a student being a reminder that they do not thereby become a different person. This makes the issue in hand less a question of how to classify successive stages in a lifecycle, and more one of dealing with the kind of change involved when one thing – the myoblast is an embryonic form of muscle cell – changes into something else.

Resuming hierarchisation, correcting mistakes

The group reconvenes for a second meeting (after several months). In the interim, a substantial amount of work has been done by the pairs proposed above, and more than two hundred cells have now been defined according to various properties. The goals of this second meeting are articulated as:

> 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 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

The sociology of science is inclined to treat the lack of stability in scientific work as its discovery - something that would be counterintuitive to scientists themselves if they were to recognise it. What is evident, we think, in the work we are describing, is that the contingent nature of the work is routinely recognised, the validity of statements about logical properties is explicitly adjudicated against institutional and professional purposes, and questions are not treated as settled except insofar as they meet the specific purposes at hand. Put simply, doing logic-in-practice involves exhibiting exactly the kind of routine corrective work that is treated only ironically in some versions of STS.

The group (the membership of which has been increased by two members who bring specific expertise) begins by looking at contractile cells (cells which contract, such as muscle cells), information about which has been gathered by one of the group members. The work being done here is that of setting out the hierarchy that was initially missing. 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:

P5: *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 P2.*

P2: *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.*

P2: *I think that is ... I think there aren't many ... but there's at least one*

P3: *the second thing is that we need some synonyms... cell biologists don't talk about fast muscle cells, they talk about muscle fibres or myofibres ...*

P1: *I don't know how rich the OBO version of the cell type ontology was but the OBO format has a mechanism separate from the textual definitions for doing various forms of synonym which if they're there, will just be transferred over but you are entirely right*

Of course, this process also entails the routine identification and correction of mistakes. Sometimes, they are easily agreed and rectified but not always, for 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 have been made and agreed:

P2: *could we just look at all the children of contractile cells?*

P5: *[runs reasoner]*.

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!!*

P6: *a flight muscle cell is never a cardiac muscle cell*

P7: *it's a sib of skeletal it's got that right it's just a contractile cell, which is right ...*

P2: *but the display looks wrong ...*

P7: *[goes to board, points fingers to each term]*

P8: *we're looking at cardiac, the highlighted one*

P7: *oh, we agree that's wrong*

P2: *so that's a good one to look at if it's wrong*

Corrective work is the main part of what is done at this late stage. As classification decisions evolve what was once 'right' may now need revision; original assumptions may have been entirely wrong; there may be sins of omission, or poor or careless input work (which nevertheless impacts on the capacity of the reasoner to function). In any event corrective work is done by those who know and know how:

P6: *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, so it develops into, it develops into, I can give you the reference for this*

P1: *how have we got it axiomatically described?*

P5: *yeah, its '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 [on screen is description of pericyte with GO IDs]…*

Exigencies of Logical and Semantic Work

We have shown how the articulation of semantic judgements and logical formalisations are mutually elaborated in two stages of a workshop designed to try out a relatively unused method, normalisation, in the design of an ontology. The work in hand is exploratory, and the first phase very much involves provisional moves as those present try to work out, often conjointly, how they are to proceed and to begin identifying candidate structures for the ontology they aim to (re)build, already attentive to the potential such structures have for both logical expansion and for adding to or alleviating the eventual burden of work. In the later phase of the work, a partially developed structure is in place, one that can be assessed to see if it has turned out to be the one that they were intending to design. The kinds of semantic issues arising relate, of course, to the order of the work-in-

hand, different questions arising in the earlier phase, when decisions are being made about how to identify and order axes for the ontology (and where the live issue was whether any effective axes could be identified or whether an alternative, more laborious method of developing the classification 'through restrictions' was to be used), to those arising later when working through the consequences of having adopted certain axes.

It is quite characteristic of this work that stable, complete and unchallenged definitions of classes, properties and relations in the ontology in question are consciously and deliberately postponed, since the structure being build is a network of interdependencies, and the finalisation of one decision often awaits fixing of other decisions[5].

This called for questions about what terms would form the metalanguage in which the revised ontology might express the domain terms, and whether the transformation of the overall structure would change the meaning of or displace terms from the original's metalanguage.

We have tried to show how practical semantic and logical work is attendant upon the various problems confronted at different times in the ontology building process, and in so doing at least intimate some of the dimensions of the ontology building task (*at this stage* of the builders' work) through displaying some of the steps involved in articulating biological terminologies with the requirements of the ontology's developing structure. One thing which needs to be emphasised is that much of what is going on is that formal ontology building rules (such as those of OntoClean) are not understood to supply automatic determinations of how some item of terminology is to be *correctly* classified, treating such determinations as matters for decision by the users of the tools in their work as designers of an ontology. Since the work on this occasion is in reworking an ontology, many of the questions addressed have to do with understanding the principles that the CTO ontology had employed and with whether these were to be preserved or modified in the reconstruction.

This work, then, can be seen as confronting a series of quite practical problems which even experts must confront on a quite routine basis. These are not by any means all ones of formal logic in the sense that there are failures in their understanding of first order logic's principles, but problems in the sense that logical procedures require implementation and instantiation. These are dealt with in an ordered fashion. That is, at each point, members of the group identify what their problem is, whether it is a problem of structure, of semantics, or some combination of the two, what they are to do about it at this time and in this case and what their rationale might be for making these decisions.

In the mutually elaborating nature of this logical and semantic work, one thing is clear – the work is the work of classification. Classification can be treated in a somewhat trivial way in the

[5] We've discussed elsewhere the work of 'elicitation' in which ontology builders engage to help in the initial identification of the vocabulary to be included in the ontology (see Lin et al., 2007), something which was not necessary for this exercise, since the CTO already provides that, but this did not and does not eliminate the need for decisions as to what biological terms should be included in the ontology, and the task of reconfiguring the overall structure of the cell ontology.

literature, exemplifying as we have suggested the 'discovery' of instabilities or (see Bowker and Star, 1999) used contrastively to show how classifications both have formal properties and entail work to produce them.

What is seldom pointed out is that this work has no generic features – classification is not done independently of the conditions of its production. There is a world of difference between a classification system for mental illness, such as the DSM-IV, and the work that is being done in ontology production. It has been remarked on a number of occasions that the classification of mental illnesses can entail vagueness, overlap and confusion (see, e.g., Healy, 2002; Lane, 2007; Kirsch, 2010). It is nevertheless commonly used by practitioners for diagnostic purposes. In contrast, the work of ontology production is *precise* work. It cannot be anything else because anything less than worked through classifications will merely cause the mechanised reasoner to signal errors. Now this does not mean that issues must be resolved, definitions must be universally agreed, and so on. It means instead that the work must be done in precise relation to the categories inherent in any ontology – those of instance, class, property, relationship, value, and so on. The work, in other words, entails orienting domain knowledge – in this case, knowledge of cells and their characteristics – into the language of formal logic.

What is noticeable in our data is how P1, the leader of the group, is extensively the authority who adjudicates questions about the proper logical form, usually by supplying illustrations to explain how these things are to be modelled. The parties have some familiarity with the ontological scheme but not necessarily of the kind that lets them enter into the work directly. They need some reminding of how the categorical scheme works, of prominent and relevant features of classification – invocation of upper level ontologies 'à la BFO' (Basic Formal Ontology), which instantiate a couple of stock problems in deciding how to enter things – particularly those of identifying inherent properties, part-whole relations and so on. Members of the group (to a varying degree) are familiar with various types of cell and how those cells are typically described in the world of biology. What they are engaged in, however, is the transformation of these typicalities into statements which are *ontological* statements.

This is not just a matter of deciding how to make entries into a formed ontology, but on how to form the axes of the redesigned ontology, so that at this juncture, the queries are not about which existing category is this instance to be assigned to but what kind of category would be needed to provide space such that this (and other expectable) case(s) can be included with the scheme.

This is clear in the way in which the meeting participants oriented, guided by P1, to the principle of 'rigidity' in the first instance as a means to begin the structuring process. 'Rigidity' is about differentiating properties which are 'essential' to the identity of some object and properties which are not essential, a difference made in the Ontoclean vocabulary which contrasts rigid with anti-rigid characteristics. Parties are nevertheless alert to the decisional status of their categories: each cell classification is a design decision of their ontology, and it can

be recognised that alternative decisions are possible. Such decisions are, of course, not arbitrary but are made in accordance with assumptions about what the problems are and who is best suited to solving them. These are innumerable negotiated outcomes.

These matters of course involve variations across the participants, broadly about difference between knowledge of ontology-principles or about biological phenomena, both of which are themselves unevenly distributed: P1 is an ontology specialist and is something of an authority on the general task, as well as on the rules of ontology building. There are varying degrees of knowledgeability in these matters, but there is also knowledgeability about biology generally and about cell types specifically, so a lot of these are queries to which there is a ready answer. Some, however, are queries which may have different answers, and yet others are queries that can't be answered now or yet.

Conclusion

Our study deals only with a few brief instances from a workshop that involved four full days of meetings as well as a practically unquantifiable amount of additional work, which workshop is only a small part of quite long term efforts at the development of an online technology. These brief moments are dense with specific understandings of a host of practicalities, and the materials of a plurality of disciplines which are somewhat unevenly distributed amongst the participants, though not in ways which create notable difficulties amongst them, the more worrisome troubles, as we have suggested, residing in the ways which what they 'know' articulates with the understandings of those who are not present or are known only as imaginable social types. These moments give some sense of the intense, dense and protracted nature of the work going into the preparation of a computational infrastructure that may potentially enable the transformation of discovering work in a science like biology through enabling this to take place on line.

References

Bietz, Matthew J., Eric P.S. Baumer and Charlotte P. Lee (2010), "Synergizing in cyberinfrastructure development", *Computer Supported Cooperative Work*, 19, pp. 245-81.

Bowker, Geoffrey C. and Susan Leigh Star (1999), *Sorting Things Out*: *Classification and Its Consequences*. Cambridge, MA: MIT Press.

Guarino, Nicola and Welty, Chris (2002), "Evaluating ontological decisions with OntoClean", *Communications of the ACM*, 45, 2. ACM. New York.

Healy, David (2002), *The Creation of Psychopharmacology*. Cambridge, MA: Harvard University Press.

Kirsch, Irving (2010), *The Emporer's New Drugs: Exploding the Antidepressant Myth*. Basic Books.

Lane, Christopher J. (2007), *Shyness: how normal behavior became a sickness*. Yale University Press.

Lee, Charlotte P., Dourish, Paul and G. Mark (2006), "The human infrastructure of cyberinfrastructure", *Proceesings of ACM Conference on Computer-Supported Cooperative Work CSCW 2006* (Banff, Alberta), pp. 483-492.

Lynch, Michael (2008), "Ontography: Investigating the Production of Things, Deflating Ontology", Oxford Ontologies Workshop, Saïd Business School, Oxford University, unpublished mss.

Randall, Dave (2001), "Review of Bowker and Star (1999)", *Computer Supported Cooperative Work*, pp. 147-53.

Randall, Dave, Procter, Rob, Lin, Yuwei, Poschen, Meik, Sharrock, Wes and Robert Stevens (2011), "Distributed ontology building as practical work", *International Journal of Human-Computer Studies*, Vol. 69, Issue 4, pp. 220-233.

Ribes, David and Charlotte P. Lee (2010), "Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories", *Computer Supported Cooperative Work* , 19, pp. 231–244.

Ribes, David and Thomas A. Finholt (2007), "Tension across the scales: planning cyberinfrastructure for the long term", *Proceedings of GROUP 07*, pp. 229-238.

Sormani, Philippe (this issue) "The Jubilatory *YES!* On the Instant Appraisal of an Experimental Finding", pp. 59-77.