White paper

Durham Zoo: prior art and solution search

Arthur Absalom, Geoffrey Absalom
durhamzoo@ymail.com

April 1, 2012

Abstract

Durham Zoo is a design proposal for a collaborative-classification-based database and search engine for retrieving both the prior art, and solutions to problems.

The design is simple and intuitive enough to empower input from the crowd. The inputted data can then be searched with a concept search engine.

The design can import classification information from existing classification schemes, such as those operating in the different patent offices around the world, in a simple way. The combined collection of classification information is then searchable in one go via the Durham Zoo codes.

Increased levels of computer intelligence could be incorporated into the design, progressing toward a combined classification-and-text search.

We believe that a global authority, similar to the Wikimedia Foundation, could develop and operate the system for the benefit of all. The IEEE would appear to be an ideal candidate for this role.

The paper is in two parts. The first part relates to the need for the system and to the broad design. For those that so desire, the second part provides supporting argument for the need, and more information as to functionality and operation of the system.

We are not academics and this is not an academic paper. It has no peer review. It is intended as a proposal to kick-start a project.

Part 1: The need and the design

Contents

1. Introduction

Prior art searching is often not as efficient a process, as all encompassing in scope, or as good in result, as it could and probably should be.

The prior art includes library collections, journals, conference proceedings and everything else that has been written, drawn, spoken or made public in any way. Much technical information is only published in patents.

There are many good reasons to improve prior art searching.

Research, industry, and indeed humanity, face the spectre of patent thickets: an impenetrable patent space that effectively hinders development rather than promoting it. Improved prior-art searching would help with the gardening and result in fewer and higher-quality patents. Poor-quality patents can reward patenting activity *per se*, which is not what the system was designed for.

Improved prior-art searching could also result in less duplication in research, and/or improved collaboration.

And we could also make much better use of the prior art.

How about designing a prior-art searching system that also supports innovation and problem solving: one that is designed to find solutions from non-obvious areas of technology or science?

What about a prior-art system that ‘pings’ a researcher when a potentially interesting document is published somewhere on the globe? It could be that the document relates to their specialist field, or maybe it presents a solution to a problem from outside their area of expertise. This is Durham Zoo Solution Search.

Whether searching the prior art, or whether searching a problem to a solution, we need a tool that searches concepts. We believe that at the moment, the only way of being able to *consistently* perform an *efficient* concept search whilst *guaranteeing* the quality of the result is with the support of classification performed by humans. It’s just we haven’t yet got the classification scheme, or the database, that we need.

But we have the Internet, which has opened up the world to ‘crowd source’ collaborative efforts such as Wikipedia. And so although a big task, we have the (people)power, and we have the technology.

Collaborative classification is not a new idea.

Collaborative Classification of Growing Collections with Evolving Facets, published by Harris Wu et al. in 2007 [1] describes a ‘wiki-like classification scheme that is similar in many ways to our design.

The European Patent Office evoked the possibility of a Web 2.0 “wiki-class” system for patent professionals at the PATLIB Conference in 2010 [2].

We believe our design can empower the crowd to input concept data in an accurate way. The means to input concept data is also the heart of a ‘fuzzy’ concept search engine, which can exploit the power of contemporary computing to search the database and get a same, or a similar, concept out.

We believe this humans-data-in and computing-data-out synergy is what we need start now. However computer intelligence is ever developing, bringing improvements in automatic classification, language translation and natural language comprehension. The semantic web is on the way.

And so whilst we believe that the Durham Zoo design could do the job now, we see it as the human interface to a future system incorporating more advanced computer science alongside the crowd's collective intellect.

2. Searching concepts: the who, the where, and the how

2.1 The who

There is the searching of information *per se*, and then there is the searching of concepts: searching a train timetable is quite different to searching a solution to a problem.

To search concepts we have human intelligence and we have computer intelligence.

Computer intelligence is on a roll. There were many who were surprised by the victory of IBM's Deep Blue over chess grand master and world champion Garry Kasparov in 1997. Many had predicted it could never happen [3].

But chess is essentially a sequence of patterns, and computers are very good at processing large sets of such data. A computer can afford a sledgehammer approach to cracking a chess nut.

In computer science, natural language processing (NLP) is the means by which a computer understands text as a human would. 'A metal can', has two possible meanings: a can made of metal, or the ability of a metal to do something. NLP can interpret the meaning from the context. And NLP is a pre-cursor to foreign-language translation: so called machine translation.

Search engines such as Bing and Google process a search query and then retrieve and rank links to potentially pertinent information sources. WolframAlpha, the computational knowledge engine, returns an actual answer to a query.

IBM was in the news again recently when Watson outsmarted two human contestants in a televised quiz show called 'Jeopardy', thus demonstrating an improved ability to understand language [4].

Siri (Speech Interpretation and Recognition Interface) is an intelligent personal assistant and knowledge navigator operating on Apple iPhones [5].

But dealing with concepts is different from dealing with facts. What of a fluffy concept, dressed in the ambiguous, synonymous, polysemous and generally flawed approximation that is human language? And then maybe further obfuscated in technology- and/or patent-speak? That would be a 'big ask' for a Turing test [6].

Watson, Siri and computer science notwithstanding, the human brain is the state of the art as concerns understanding both language and concepts.

Humans have long been 'classifying' disclosures with different shorthands to help identify the content of a disclosure.

Such systems are in use in the world's patent offices, with many classification schemes linked to a greater or lesser degree to the International Patent Classification (IPC). Publishers and information providers also operate a wide range of proprietary classification schemes, keywords, indexes and tags.

Computers can also be used to classify or 'index' disclosures. Such classification can be to an existing scheme such as the IPC [7], or to a custom-designed scheme [8].

However, given that computers cannot understand a concept to the same degree as a human expert, so a computer cannot classify with the same accuracy or degree of granularity.

Support for this argument comes from Galaxy Zoo, the project to classify the galaxies in our universe [9].

Galaxy Zoo uses the human brain to classify the structure of the galaxies in a web-based collaborative effort. The brain is ‘better than the most advanced supercomputer’ in image processing and classification. Interestingly the image processing and classification doesn’t include a human language factor.

The ‘crowd’ in the crowdsource collaboration solves the processing scalability problem.

And yes, Durham Zoo has taken the Zoo from Galaxy Zoo. The Zoo is intended to be in homage to Galaxy Zoo rather than simply plagiaristic.

As regards concept searching, we believe a powerful solution would combine the power of the crowd with the power of contemporary computing.

Such a system needs to be simple enough for humans to add information, whilst being able to exploit the processing power of computers to make best use of the stored information.

Put simply: humans put concepts in, and computing gets concepts out.

However, given the constant evolution of computer intelligence, we should endeavour to produce something that is future-proof in design.

Web 3.0, otherwise known as the ‘semantic web’, is on the way. Web 3.0 implements the tagging of data with metadata to make information more understandable to computer software.

Tags can be grouped into ontologies that map the relationships between entities. There is already an ontology language for the Internet, called OWL. And with ontologies and natural language processing comes automatic classification.

We believe that the system we have designed could work with what we have now. And whilst Durham Zoo system could eventually be integrated with ‘semi-automatic’ or ‘computer-assisted’ classification systems, are we humans not best placed to design a scheme that computers could eventually classify to?

The end goal is to fully integrate search using Durham Zoo classification information together with ‘computer powered’ search of document text and metadata. We call this a Classification And Text search or CAT search. But that would be for later.

If ever computer intelligence truly has the upper hand and searching doesn’t need humans, we will doubtless see a reduced need for engineers and scientists, and of course the end of the patent system, given that the fictitious ‘skilled person’, by who’s knowledge an invention is judged, will have been usurped by a ‘skilled computer’.

We will finish with a demonstration of the human brain’s powers of inference: a Tommy Cooper joke: last weekend I tidied up the attic with my wife… dusty, dirty, covered in cobwebs… yeah, but she’s good with the kids.

2.2 The where

There are many repositories of information in the world, in many languages, in many formats, and with varying degrees of accessibility.

Humans and computers can only search information that is accessible to them. And even when ‘freely’ accessible the information needs to be searchable in a practical manner.

A ‘prior art’ search is performed during the examination of a patent application. Patents are public by design. Tools exist to search in patent collections with classification symbols and in the text.

Much recent prior art, is at least in part, somewhere on the Internet. But much prior art is still on paper and cannot be crawled by computers.

Much prior art is subject to copyright and thus not freely available. The costs of downloading all potentially interesting prior art identified in an Internet search of document abstracts, may be prohibitively expensive.

And then much prior art will likely be in foreign language collections.
Thus a ‘complete’ prior art search is a practical impossibility, and even a ‘thorough’ search of the prior art may be a protracted and inefficient process of uncertain result.

A globally accessible database of bibliographic information and associated classification information would facilitate a search on a very broad range of documents. The database should include an address indicating where the information could be found.

The retrieval of the pertinent documents, in their different formats, in different languages, and with different accessibility and copyright constraints would then be another task.

If a searcher can be assured of the pertinence of a disclosure, there is better sense in justifying the time, and potentially the cost, of retrieving a copy. The entire search process may thus be rendered more efficient.

2.3 The how

Whilst a prior art search may approximate to a concept search, they are not the same. From the Durham Zoo perspective, a concept search includes searching for a solution to a problem that is as yet unknown.

Searching for an unknown solution requires a search using a definition of the problem. And concept search should be able to evaluate solutions from across technology and the natural world.
We believe that many of the present systems are better designed to find the prior art than they are to find a novel solution to a problem.

We believe that the way to providing an efficient concept search, with a high quality result, is via *a classification scheme*:

i) adapted to concept searching;

ii) adapted to the crowd;

iii) adapted to contemporary computing;

iv) operating in a global database.

Our proposal is in the next section.

3. The Durham Zoo design

To tie together the requirements identified above into a working system we have identified a three-part sequence, with the results achieved at each stage:

i) we need speak the same language

> empower the crowd classification effort

ii) add 'shades of grey' classification

> automatic ranking of documents retrieved in a search
> simplification of the search process
> reduction in the classification effort

iii) add 'Multiple Aspect Classification' in five dimensions

> fuzzy concept searching

3.1 Speaking the same language: a controlled vocabulary

Language is often imprecise or ambiguous.

There are many 'variables' in language, such as synonyms, different words meaning the same thing, and polysemes, different things meant by the same word. Senders and receivers of language may have different preconceptions, different perspectives and different understandings.

Opportunities for confusion occur mercilessly Murphy-fully, as befits the law.

A controlled vocabulary is a lexicon with no synonyms and no polysemes.
Each word means one thing; each thing is described by one word only.

Perhaps the best example of a controlled vocabulary is in chemistry, a science that benefits from a universally accepted naming of the chemical elements.

We need design a controlled vocabulary for the whole of technology. Each entry in our controlled vocabulary will be a 'zootag'.

Each zootag includes an associated text file explaining the Durham Zoo interpretation of the technology and its terminology. There is thus a 'controlled meaning'.

Any synonyms that may be used to define the zootag entity are included in the associated text: they are used in Durham Zoo to help a user navigate to the controlled vocabulary.

Different zootags may have the same root or suffix, but they are all distinct. Zootags may define single entities such as chlorine, or group entities like the halogens.

There will doubtless be differences of opinion in the creation of a controlled vocabulary for technology. However classification is about pragmatism and not academic rigour: classification values consistency above the search for an absolute truth.

A protocol to achieve the necessary consistency is presented in Part 2.

Zootag 'steering' is the means by which different classifiers, with different disclosures described in different terminology, but presenting a same concept, arrive at the same zootag end point.

Each zootag, defining each entry in the controlled vocabulary has its own zootag steering diagram (ZSD). Each ZSD is a graphical network of related entities. A ZSD is in fact a simple ontology.

An example ZSD is shown below. It is the ZSD for a horse zootag.

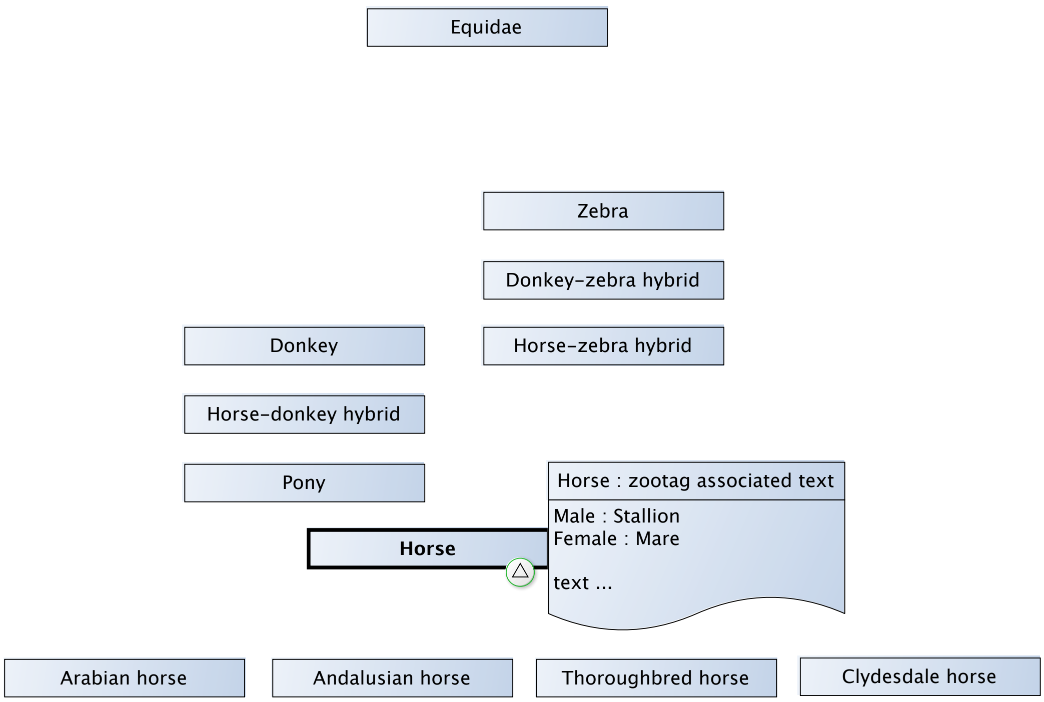

Fig. 1. ZSD for a horse

As can be seen, the horse entry, the ZSD Subject, sits highlighted in the middle.

Above the horse are entities that are similar to a horse, referred to in Durham Zoo as Inferrands, whilst below the horse are specific Examples of a horse.

A classifier can move the cursor above any of the entries in the ZSD and the associated explanatory text will appear.

Thus a classifier can surf around the information in a ZSD to discover the controlled vocabulary and the Durham Zoo interpretation of its meaning.

An indication of the similarity of the related entities with respect to the horse is apparent from their vertical position on the ZSD. So a pony is more like a horse, and thus closer on the diagram to the horse, than a zebra.

If a classifier decides that a pony is 'more like' the entity they are looking for, they can click on the pony entity on the horse ZSD to load the pony ZSD.
Each entry in the ZSD is in fact a zootag with its own ZSD.

There are divergent opinions as to what actually constitutes a pony. The pony ZSD provides the information as to the diverging opinions, includes the zootag interpretation of things to be used for classification, and includes pointers to zootag alternatives.

The zootag steering can proceed in a 'getting warmer - getting warmer' manner until the classifier knows that they are at the right zootag.

As an example, our classifier has to classify the entity that is a hybrid animal: a cross between a female donkey and a male zebra. The classifier knows that there is likely to be alternative terminology for a female donkey crossing with a male zebra, and a male donkey crossing with a female zebra.

The classifier is unsure whether there is a commonly accepted term. The classifier begins by searching for an Entry ZSD.

The classifier knows that a mule is a cross between a donkey and a horse, but decides to search in the zootag database with 'donkey'.

Both the zootag controlled vocabulary is searched as well as the zootag associated texts which contain synonyms and alternative spellings.

This Entry ZSD search may result in multiple hits. In this case, in a manner analogous to the disambiguation of Wikipedia, the best guess is selected.

In our example the donkey ZSD is identified and selected.

The donkey ZSD is shown below.

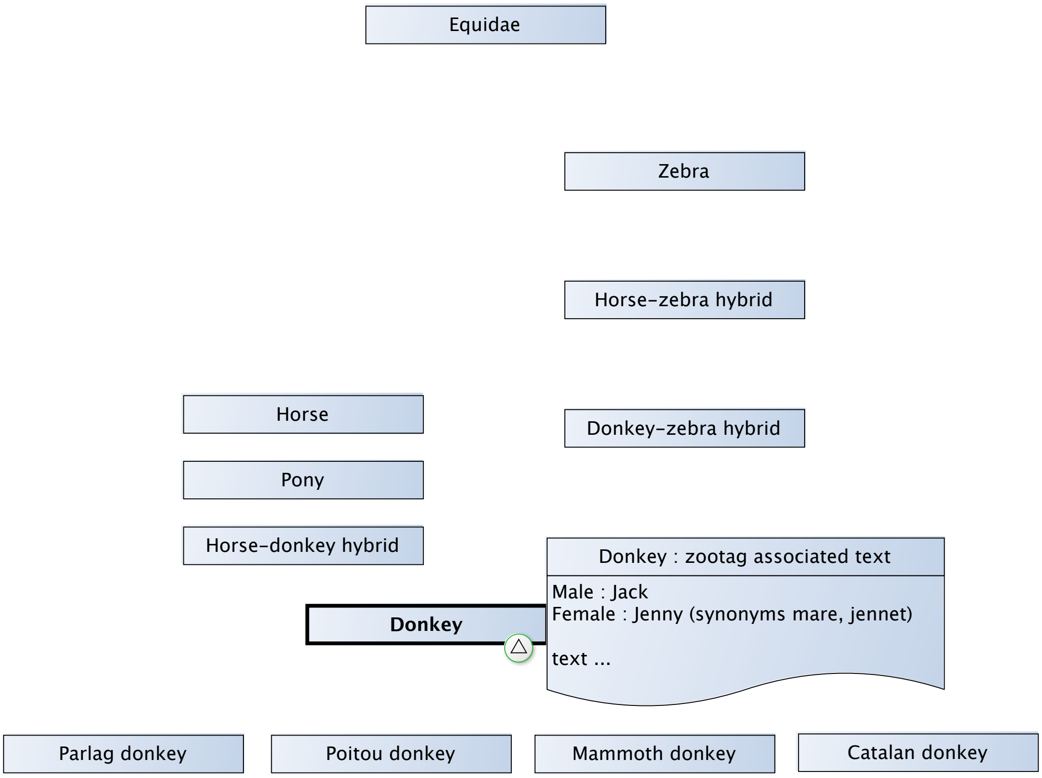

Fig. 2 ZSD for a donkey

The classifier identifies the zootag for the donkey zebra hybrid and clicks on it. This loads the ZSD of the donkey zebra hybrid, revealing more specific and more pertinent information. Only the bottom part of the diagram is shown in the figure below.

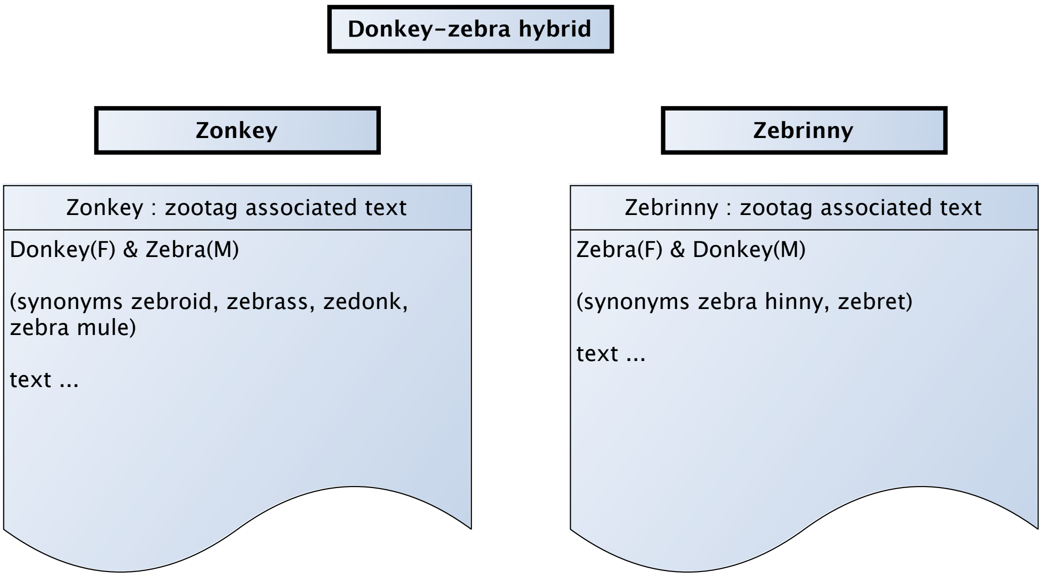

Fig. 3. ZSD for a Donkey-zebra hybrid

And so the classifier can see that the offspring of a male zebra and a female donkey is called a zonkey in the Durham Zoo controlled vocabulary, and that zebroid, zebrass, zedonk and zebra mule are synonyms.

The offspring of a female zebra with a male donkey is called a zebrinny in the Durham Zoo controlled vocabulary, with zebra hinny and zebret as synonyms.

The example above is a simple one. The presentation of related entities together in a graphic format, where the degree of similarity is readily apparent, is designed to aid both the classifier and searcher to find what they need.

The classification process may be more direct, it may be more convoluted; but either way the end point is the same.

3.2 Shades-of-grey

Much classification is yes or no. In the days where disclosures were uniquely on paper, such classification resulted from the necessity of a disclosure being put on a shelf, or not.

Such a 'black-and-white' approach does not mirror information, which is better described in terms of 'shades-of-grey'.

Is Pluto a planet? Officially no: not any more. But the matter was of academic debate and indeed controversy before Pluto was demoted from its planet status to that of a dwarf planet. There is still much divergent opinion and there are continued attempts to reinstate it as a planet.

How to classify such information?

The world of 'on the shelf or not', of '1 or 0', is the binary world of Boolean logic. In the Boolean world would Pluto be classified as a planet or as a dwarf planet?

But classification is performed for search, and so we should maybe better ask ourselves how a searcher would look for Pluto in a library of two groups: planet and dwarf planet?

The consequence of classifying in only one group could result in a searcher searching the 'wrong' group and 'missing' Pluto.

To be sure of not missing Pluto the searcher would have to search both groups. This may not be efficient.

Pragmatists may avoid this problem by advocating a double classification. The search of either group would then find Pluto.

However too much pragmatism of this kind may result in the groups becoming bloated: a dwarf planet in the planet group and a planet in the dwarf planet group may be considered as 'noise' rather than information.

If the double classification were not systematic, the risk would be of having to search both groups anyway, this to be sure the search was complete. And that could result in seeing many documents twice.

If the number of 'Pluto-like' planets were to grow the continued separation of the groups could be called into question.

The Pluto example is perhaps too simple: anyone doubting the shortcomings of a Boolean classification of information should refer to the nomenclature of Ceres the dwarf planet.

Ceres is in the asteroid belt. However in Greece, Ceres is called Demeter, which in English usage is the name of an asteroid, notably 1108 Demeter, which is in the asteroid belt along with Ceres. Between 1955 and 1975 Demeter was also a name *sometimes* given to Lysithea the satellite of Jupiter. And Demeter was also a name given to a French micro-satellite.

Sometimes information is not easily put in boxes: there are too many ‘it depends’ as with the time and language variables in the Ceres example. Things may be something to a certain degree, rather than a simple yes or no.

We know that if the groups are stored in a computer we can search the groups using Boolean operators: an AND will identify those documents that are classified in both groups for example. And we know that searching in the document text is perhaps another option.

We know that a search for ‘Ceres’ in combination with an astronomical classification code would exclude Demeter the cat from Andrew Lloyd’s musical ‘Cats’ and Demeter the fictional Russian ship that brought Dracula to England. But classification codes themselves can be smarter.

No, Boolean and black-and-white doesn't reflect the real world. Information is more like the world of quantum mechanics than classical physics: it is a world of probability and inference.

3.2.1 Shades-of-grey classification

On the horse ZSD anything appearing below the horse is a specific Example of a horse and thus fully a horse. Anything above the horse is not a horse but the closer it appears to the horse in the vertical plane, so the closer it approximates to a horse.

Thus the position of the different entities on the ZSD represents a sliding scale of likeness. The ZSD has been edited to include a ‘likeness factor’ on the right hand side.

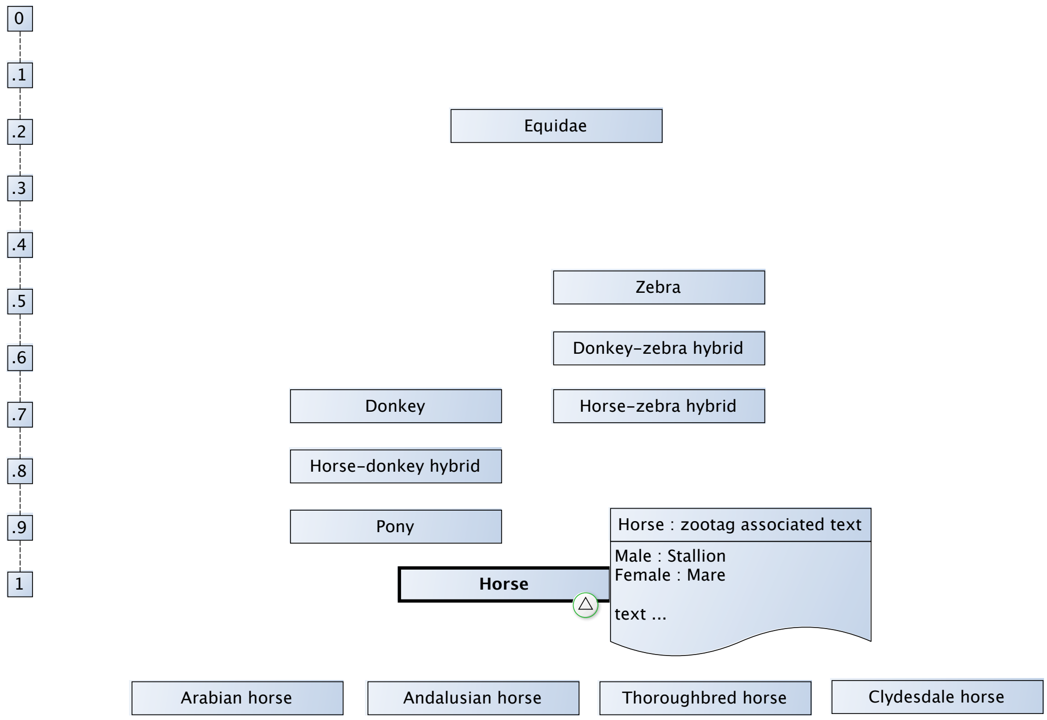

Fig. 4. ZSD for a Horse zootag with inference scale

The classification of a disclosure about a horse with the horse zootag classifies the disclosure fully as a horse disclosure on the horse ZSD.

At the same time, it also classifies the same disclosure as something similar to a donkey on the donkey ZSD, and to the appropriate degree of similarity everywhere else that the horse zootag appears on a ZSD.

The ‘something like’ represents the shades-of-grey classification.
By virtue of the library of ZSD’s the addition of a single zootag can result in multiple shades-of-grey classifications for different entities.

The ZSD-based classification is thus very efficient in terms of classification effort.

3.2.2 Shades-of-grey search

Now from a mathematician’s perspective, a search is not complete until every document in existence has been examined.

For the engineer and pragmatist, a search should start with the documents most likely to be pertinent, and continue down a ranked list until such time as the chances of finding a better document are minimal.

In Durham Zoo, whenever a zootag is selected for input into a search engine, all the entities on the ZSD are selected, each weighted according to their degree of likeness.

In a search for a horse, documents zootagged with the horse zootag, and any documents tagged with zootags that are Examples of the horse zootag, will be ranked the highest.

After that will be documents zootagged with the pony zootag and so on up for all the entities on the horse ZSD.

The shades-of-grey classification can be seen as the means of effectively ranking pertinent documents in search.

Searching with a complete ZSD, rather than a single entity is also a very efficient way to search: a searcher defines the specific, but everything that is related is also automatically incorporated into the search process, and this to the degree that each related entity is pertinent.

3.3 One big MAC and fuzzy concept searching

Multiple Aspect Classification, or MAC, is the process of classifying a concept with multiple classes, where each class belongs to an independent family of classes relating to a different aspect.
We believe that a MAC approach is the best way of defining a concept in shorthand.

We also believe that commonality should be promoted across the whole of technology to enable very different entities to be classified with aspects that are the same.

And we need a structure that is simple enough for us humans to add information, whilst able to exploit the power of computing to get good results in a concept search.

And so we have created a 5 dimensional MAC for our zootag-controlled vocabulary.
Any zootag must exist in one of the five dimensions.

The five dimensions can pinpoint a concept in a manner analogous to triangulation in navigation.

The dimensions correspond to questions that may reasonably be asked to define a concept, namely: what, where, when, why and how. More particularly:

what technology are we talking about?
where is the technology applied: is it for a particular application?
when is the concept pertinent: is there a particular time or operating mode?
why do we have to develop the concept: i.e. what is the problem?
how do we overcome the problem: i.e. how is the solution realized?

As an example, the use of a simulated sharkskin surface to reduce friction on a ship's hull:

t - technology	surface material
a - application	ship's hull
o - operating mode	operational use*
p - problem	drag
s - solution	simulated sharkskin

(* i.e. as opposed to the design, manufacture, decommissioning or recycling)

In a search in the Durham Zoo database for such a concept, the most pertinent documents would be those with a full weighting in all of the five dimensions.

However it may also be necessary to consider those documents with concepts that approximate to the searched subject matter. The use of the ZSD's of the five dimensions, each in shades of grey, allows a 'fuzzy' calculation of documents that approach the searched concept.
There are two simple methods.

The first is to calculate and thus rank documents according to a simple summation of the inference values in the five dimensions.

The second is to rank documents according to their 'vector distance' from the searched concept.

For the sake of simplicity, we will consider an example where only three of the dimensions have been used to define the concept. The three dimensions may be thought of as being in the X, Y and Z planes.

The concept we are searching may be represented as (1, 1, 1). Two documents have been retrieved with inferences (1, 1, 0) and (2/3, 2/3, 2/3).

Using the simple summation method the two documents are equally pertinent given that they both sum to 2.

However using the a vector distance calculation the first document is distance 1 away from the target, whilst the second document is $\sqrt{1/3}$ away (using Pythagoras) and thus considerably closer.

The vector distance calculation is the best method of identifying the relative similarity of documents in a MAC database.

The vector calculations can be performed in up to five dimensions using a generalisation of the Pythagorean theorem [10].

$$d = \sqrt{\left(1-t\right)^2 + \left(1-a\right)^2 + \left(1-o\right)^2 + \left(1-p\right)^2 + (1-s)^2}$$

Where a dimension is not to be taken into consideration, the term needs to be removed.

The ranked results would be dependent on the dimensions chosen to define the concept, the chosen calculation method, and of course the documents in the database.

The accuracy of the method is also dependent on the how accurately the zootags represent the concepts they define in the documents in the Durham Zoo database. And this in turn is a function of how accurately the individual zootags represent the concept they are being used to define: called the zootag variation, and the accuracy of the zootagging, the zootagging variation.

In our simulated sharkskin solution example, it could be that a document about swimming trunks with a simulated sharkskin surface to reduce drag would be ranked very highly.
If the technique had only ever been applied to swimming trunks, i.e. not known in the context of ship's hull design, the swimming trunk document could be the most pertinent.

A document relating to the problem of the fouling of ship's hulls, and thus not zootagged with a drag problem zootag, would likely be picked up and ranked highly given that the drag zootag would include the fouling zootag as a closely related entity on its ZSD.

The search process is thus both efficient and powerful.

Furthermore, the ZSD's, which are the means for steering the classifier to the controlled vocabulary, form the basis of the fuzzy search engine.

More information about the design and operation of Durham Zoo search engine appears in Part 2.

Suffice to say here that we foresee the development of 'search power' as the DZ database grows and the data therein is further refined, and as the search engine evolves to incorporate additional functionality such as NLP.

4. Supporting innovation: Durham Zoo Solution Search

Patent information is a good place to find solutions to problems. Eighty percent of information in patents is uniquely published in the patent. Ninety percent of patented solutions are no longer subject to patent protection.
Patent literature is thus a very rich source of information and ‘free’ solutions.

A solution to a problem may be found searching the prior art with classification codes, with a text search, or a combination of both.

Many patents are classified using the International Patent Classification (IPC) scheme, or a derivative of it. Although there are IPC classes relating to general technology, the inclusion of an IPC class may restrict the search, and thus any solutions, to a particular technical field.

And how much information from particular application fields finds its way into general technology groups?

We believe that the Durham Zoo design, that classifies problems and solutions with greater commonality, will improve the chances of discovering solutions to problems from ‘unexpected’ technical fields.

This extends the functionality of a prior-art searching tool into the domain of innovation management. Innovation management includes processes such as innovation creation through analogy [11], and looking to the natural world for inspiration: known as biomimetics.

Going back to the sharkskin example. Many ship designers have long known that a sharkskin covering on the hull of a boat can have two beneficial effects: the drag through the water is reduced, and the hull resists the attraction of barnacles and other foreign bodies, thus eliminating the periodic scrubbing of a fouled hull in a dry dock [12].

However when we started Durham Zoo, we did not know if artificial sharkskin was used in the design of stents, the artificial tubes that can be inserted into the plumbing of the human body (and this was the solution to ‘our’ problem).

Stents have similar problems to ship’s hulls: they can become fouled and subsequently clog, necessitating their replacement, which can have serious repercussions.

A biomedical engineer looking for a solution to the problem of clogging stents may not instinctively look in the field of foul-prevention of ships hulls.

We searched the prior art to see if such a solution was known. We surfed the Internet and found many results for ‘stent and sharkskin’. Most of the search-engine-obtained hits contained stent and sharkskin in completely separate parts of a same web page or document, the two aspects effectively unconnected.

We found leads to follow up by searching in German including a reference to ‘Hai Tech’ (Hai is shark in German) [13].

We found literature describing sharkskin’s anti-bacterial qualities, and a synthetic imitation for use in a catheter [14].

Our best result: WO01/80919, makes a reference to stent design and the ‘favourable flow behaviour to liquids (e.g. shark skin or lotus effect)’ [15].

However importantly, our best result was found using a search that included sharkskin in the query. In this respect it was more a search of the ‘prior art’ than a ‘solution search’. More information is included in Annex 1.

The thin reference to sharkskin in the patent may result from the idea being more generally known in stent design. Alternatively the reference could be a ‘Fermat’s last theory’ like remark, we don’t know. Either way, there was a time when the link between sharkskin and stents had not been made.

We believe that the definition of concepts in five dimensions in Durham Zoo would use the prior art to better effect, and help make such links.

And if a researcher could define and ‘post’ a problem in Durham Zoo, they could receive notification of the classification of a disclosure with a potentially interesting solution to it.

5. Let's Work Together

A Durham Zoo-like system has the potential to become a global technological library. As well as helping the patent system and supporting industry and academia, the project could make money.

Targeted advertising could be implemented: during the classification and search processes advertisements related to the subject matter could be displayed.

As an example, a person searching for information on a carbon-framed bicycle may be keen to receive details of such systems for sale. If they were, then advertisers would doubtless be prepared to pay.

Any monies generated could finance the project, be returned to the classifiers, could fund research and academia, or generate profit. We don't think the last option is appropriate: we believe it should be 'by the people, for the people' to quote Abraham Lincoln out of context.

We also need a system that scrupulously respects copyright. Any trivializing of copyright to generate revenue through targeted advertising should not be undertaken.

But the crowd will hopefully do the work and the crowd should ultimately decide.

We would advocate a small-scale pilot in two technical fields, perhaps error coding and traffic cones?

Error coding is theoretical, mathematical, maybe suited to taxonomy and computer programmers?

Traffic cone technology is a more general technology, has perhaps a broader church, and perhaps a special place in popular culture: for example there is the Traffic Cone Preservation Society [16], and there is the Guinness World Record traffic cone collection [17].

We could sponsor a Durham Zoo's Cone Crazy initiative?

Traffic cone technology is surprisingly diverse: there are patent applications relating to their manufacture, their storage, their cleaning, their deformation in case of accident, and at least one for a legion of motorized and GPS controlled traffic cones to go from configuration A to configuration B without human assistance.

It would appear reasonable to look for partners to the project from industry, academia and the patent offices. Mozilla has a 'prior art' initiative; academia can contribute both with computer science and innovation management; and the patent offices have much expertise in classification and enormous collections of classified documents.

Finally, whilst searching sharkskin biomimicry we have come across the TRIZ project.

TRIZ is a methodology for inventive problem solving. A set of inventive principles has been identified following an extensive study of patented inventions. The methodology involves an analysis of the often-contradictory consequences of implementing a potential solution.

We would suggest that a further investigation of TRIZ be undertaken as a first step, this to identify opportunities for collaboration or the integration of some of the TRIZ DNA into Durham Zoo [18] [19].

We will sign off with the inspirational words of Canned Heat, the blues-rock band that appeared at the Woodstock Festival in the 1960's: 'Let's Work Together' [20].

6. About the authors

Arthur Absalom

On leaving Rugby School in 2007 Arthur matriculated at University College, Oxford to read Engineering Science. In a busy and stimulating first year Arthur learnt that he didn't want to be an engineer. Arthur left Oxford to study French and play rugby in the Pyrénées-Atlantiques and in Paris. Still in one piece, Arthur is now reading Mathematics at Durham University.

Geoffrey Absalom

On leaving the British School in The Netherlands in 2005, Geoffrey matriculated at Pembroke College, Oxford to read Engineering Science. Geoffrey graduated in 2010. His Master's dissertation investigated the stresses in quantum dots. Geoffrey is now a high-frequency fx quant/trader in London.

7. References: Part 1
(Internet links valid 31.03.2012 (some need a ‘return’ in the browser)

1. Collaborative Classification of Growing Collections with Evolving Facets, Harris Wu et al., 2007
http://facet.cs.odu.edu/publications/HT07facetsmanchester.pdf

2. Classifying patents using Web 2.0? H. Wongel et al. PATLIB, Dresden, Germany, May 2010.
http://www.epo.org/learning-events/events/conferences/2010/pi-conference/discussion-rounds.html

3. Deep Blue
http://www.research.ibm.com/deepblue/meet/html/d.3.html

4. IBM WATSON
http://www-03.ibm.com/innovation/us/watson/index.html

5. Siri (software)
http://en.wikipedia.org/wiki/Siri_%28software%29

6. Turing test
http://en.wikipedia.org/wiki/Turing_test

7. Automated Categorization in the International Patent Classification, C.J. Fall et al., 2003
http://www.sigir.org/forum/S2003/CJF_Manuscript_sigir.pdf

8. Patent US 2007/0136221 A1, Sweeney et al. (14-06-2007)
www.google.com/patents/US20070136221

9. GALAXY ZOO
http://www.galaxyzoo.org/

10. Pythagorean theorem
http://en.wikipedia.org/wiki/Pythagorean_theorem

11. Generating innovations through analogies, K Kalogerakis et al., 2005
http://www.tu-harburg.de/tim/downloads/arbeitspapiere/Working_Paper_33.pdf

12. Projekt: Giftfreies Antifouling nach biologischem Vorbild, A. Kesel et al., 2004 (bionik hs-bremen)
http://www.hs-bremen.de/internet/de/forschung/projekte/detail/index_18235.html

13. Herstellung funktionaler Oberflächen, J. Schulze, 2009
http://www.konstruktionspraxis.vogel.de/themen/werkstoffe/oberflaechen/articles/172250/

14. current topics in biomimetics, Kerry, 2009
http://biomimetics-biol0170.blogspot.com/2009/11/preventing-bacterial-growth-in.html

15. Patent WO 0180919 A2, Grunze et al. (01-11-2001)
http://worldwide.espacenet.com/publicationDetails/biblio?CC=WO&NR=0180919&KC=&FT=E&locale=en_EP

16. Welcome to the Traffic Cone Preservation Society
http://www.trafficcone.com/

LARGEST COLLECTION OF TRAFFIC CONES
17. http://www.guinnessworldrecords.com/world-records/size/largest-collection-of-traffic-cones

18. Sharks with arms – exploring the transformation from natural phenomenon to innovation object, via applied bionics, H. Skulberg, 2009
http://ocs.sfu.ca/nordes/index.php/nordes/2009/paper/viewFile/228/171

19. Biomimetics - Borrowing the Best from Biology
http://wiki.bath.ac.uk/display/OOB/Biomimetics+-+Borrowing+the+Best+from+Biology;jsessionid=CC6AC0BFADA4841FBACC45E2831C0F82

20. Canned Heat: Let's Work Together
http://www.youtube.com/user/myBeatclub - p/search/0/oXo6G5mfmro

,

8. Annex 1: Sharkskin technology and the IPC.

Where are sharkskin technology patents classified? And is there a general technology group for reducing surface drag?

Entering ‘sharkskin drag reduction’ into Google Patents produces

US 2010/0278011
IPC Classification G01V 1/38
US Classification 367/20
SYSTEM AND METHOD FOR TOWED MARINE GEOPHYSICAL EQUIPMENT

US2005/0008495
IPC Classification B63H 1/26
US Classification 416/241
REDUCTION IN THE NOISE PRODUCED BY A ROTOR BLADE OF A WIND TURBINE

US2002/0097525
IPC Classification G11B 5/48
US Classification 360/244.2; 360/97.02
DISC DRIVE ACTUATOR ASSEMBLY DRAG REDUCTION FEATURES

US 2010/0108813
IPC Classification B64C 1/38
US Classification 244/130
PASSIVE DRAG MODIFICATION SYSTEM

US2008/0061192
IPC Classification B64C 21/10; B64C 21/00; B64C 23/00; B64C 9/00; B64C 21/06
US Classification 244/200; 244/198; 244/201; 244/205; 244/209
METHOD AND APPARATUS FOR MITIGATING TRAILING VORTEX WAKES OF LIFTING OR THRUST GENERATING BODIES

The diversity in application is perhaps testament to the widespread knowledge of the use of sharkskin to reduce drag. Interestingly a general technology group for drag reduction is not readily apparent.

The IPC groups:

B63B 1/34 SHIPS OR OTHER WATERBORNE VESSELS; EQUIPEMT FOR SHIPPING: HYDRODYNAMIC OR HYDROSTATIC FEATURES OF HULLS OR OF HYDROFOILS: OTHER MEANS FOR VARYING THE INHERENT HYDRODYNAMIC CHARACTIERISTICS OF HULLS: BY REDUCING SURFACE FRICTION

and

B64C 1/00 AEROPLANES; HELICOPTERS
FUSELAGES; CONSTRUCTIONAL FEATURES COMMON TO FUSELAGES, WINGS, STABILISING SURFACES, OR THE LIKE

are very pertinent but there are no references in the IPC for a more general technology group. Apparently there is no commonality in terms of classification of the sharkskin on a boat and the sharkskin on a rotor blade. And what of the dimples on a golf ball?

Again the JPO's viewpoint scheme may provide more information. We have not investigated it.

Part 2:
The project in more detail

Contents

1. The patent-world designs and why we need something different.

1.1 The structure of existing schemes

There are many classification schemes in existence, of differing design and granularity: the IEEE has a tagging scheme for example, and there are schemes operated and run by the IET Inspec, and Derwent's World Patent Index.

The world's patent offices have highly developed schemes based to a lesser or greater degree on the International Patent Classification (IPC).

An appreciation of the different structures has been helpful in designing Durham Zoo.

There are essentially two structures in the classification schemes:

i) a hierarchical root structure where roots relating to a common entity are sub-divided into ever-smaller roots, the information becoming more specific in nature as the roots get finer;

ii) groups of like entities grouped together in families, typically with a selection of one within the family for multiple families.

The first structure is examined by way of the IPC; the second structure by way of Japanese Patent Office viewpoints.

1.1.1 The IPC: the root *is* the problem

Paper libraries are commonly designed along the same broad lines: they are split into main sections, each of which is split up into sub-sections, which in turn are split into smaller sub-sections: for example: history, French history, post-war French history.

It is an obvious structure to classify books or paper documents. The structure, when represented graphically, is a root structure under each of the main sections.

The IPC is based on a root structure. There are eight main roots A-H that together cover the whole of technology [1].

The main roots can be viewed as the root systems of eight separate trees given that there are no links between them. That said, the classification scheme is accompanied by explanatory notes indicating on those of the other roots related prior art can be found.

The root structure is best suited where the groups are distinct and non-overlapping given that multiple classification, effectively putting multiple copies of books on multiple shelves of a paper library, is very costly. Such a cost constraint is not applicable in a paperless or electronic library.

And neither is a paperless library constrained by the three dimensions of our physical world. Documents can be ordered in as many virtual dimensions as desired, which opens up the possibility of more powerful classification structures.

Thus to restrict a paperless library to a root structure is to impose an unnecessary constraint.

Technology is always changing: a root structure is difficult to adapt to these changes. A good example is exemplified by digital convergence. Digital convergence: the increasing similarity of computing, telecommunications and television, has not been paralleled by a similar convergence in the IPC classification scheme.

In the days when a computer was typically a stand-alone device, telecommunications typically meant copper-wired telephony, and television was a cathode ray tube receiving analogue signals sent through the ether, classification across the different shelves was relatively simple.

These days a modern smartphone may include all three functionalities.

Classification theory says that a document is analysed and classified according to its inventive concept: so not everything about a smartphone need be classified in computing, telecommunications and television.

However in an increasing number of cases the classification theory, which is very simple, would appear tricky to apply. Simple is not the same as easy.

And so whilst it may be *relatively* straightforward to understand what a technical document is about; it may be increasingly difficult in some technical fields to know where it actually goes in the IPC, or where to find it.

Consider the updating of software in a smartphone. A quick search in Google Patents [2] or in the European Patent Office’s Espacenet system [3] will produce prior art with a wide range of classification codes. True, the classification classes may relate to different aspects of updating the software: from a network aspect, a security aspect, or a reliability aspect.

However computer technology is on the G root and telecommunications on the H root. The divide between the technical fields being difficult to define may result in classes on the G root and H root defining similar content.

The danger of such a situation is the creation of pockets of similar prior art that are distinct and not linked by the classification scheme. And each individual pocket risks being incomplete.

If all the pertinent groups are identified, a resulting search will at best be inefficient. If all the pertinent groups are not found, the search result may be poor.

Nanotechnology poses a different problem: whereas digital convergence is things becoming similar, nanotechnology is often multi-disciplinary.

When things get really small, when we enter the world of quantum effects, the split between chemistry, physics, biology and engineering gets fuzzified.

The IPC was not designed for such things as a single-DNA molecule nanomotor regulated by photons [4].

Elsewhere a subject that is recognised in industry as a single subject is peppered across the IPC: Reliability Engineering for example can be found in many different application fields but there is no general field. To identify all of the application fields to perform a search for a more general concept may be problematic.

Even in mature technologies, the root structure of the IPC represents a missed opportunity in the lack of any inferences across the individual roots.

Patent US 2006033674 A1 is classified at the European Patent Office according to the IPC-based ECLA scheme. The patent has classes in seven out of the eight main groups of the IPC [5].

The ECLA classes are produced twice below. The first time the technical field is highlighted in red. The second time invention-related information of the classification is highlighted in different colours.

Whilst there is obvious commonality across the different major roots of the IPC, there are no links or inferences between them.

In red, the technical field

HUMAN NECESSITIES	A
AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING	A01
ANIMAL HUSBANDRY; CARE OF BIRDS, FISHES, INSECTS; FISHING…	A01K
Culture of fish, mussels, crayfish, lobsters, sponges, pearls or the like	A01K61
Floating fish-farms	A01K61/00F
PERFORMING OPERATIONS; TRANSPORTING	B
PHYSICAL OR CHEMICAL PROCESSES OR APPARATUS IN GENERAL	B01
SEPARATION	B01D
Processes of separation using semi-permeable membranes…	B01D61
Reverse osmosis; Hyperfiltration	B01D61/02
Energy recovery	B01D61/06
CHEMISTRY; METALLURGY	C
TREATMENT OF WATER, WASTE WATER, SEWAGE, OR SLUDGE	C02
Treatment of water, waste water, or sewage	C02F1
By heating	C02F1/02
By distillation or evaporation	C02F1/04
Using solar energy	C02F1/14
MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING ENGINES OR PUMPS	F
HEATING; RANGES; VENTILATING	F24
PRODUCING OR USE OF HEAT NOT OTHERWISE PROVIDED FOR	F24J
Use of solar heat, e.g. solar heat collectors	F24J2
Solar heat collectors having working fluid conveyed through collector	F24J2/04
Having concentrating elements	F24J2/06
Having reflectors as concentrating elements	F24J2/10
Parabolic	F24J2/12
Flexible	F24J2/12C

MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING ENGINES OR PUMPS	F
HEATING; RANGES; VENTILATING	F24
PRODUCING OR USE OF HEAT NOT OTHERWISE PROVIDED FOR	F24J
Use of solar heat, e.g. solar heat collectors	F24J2
Component parts, details or accessories of solar heat collectors	F24J2/46
Arrangement of mountings or supports	F24J2/52
Airborne solar collectors, e.g. using inflated structures	F24J2/52D
PHYSICS	G
MUSICAL INSTRUMENTS; ACOUSTICS	G10
SOUND-PRODUCING	G10K
Methods or devices for transmitting, conducting or directing sound in general…	G10K11
Methods or devices for transmitting, conducting, or directing sound	G10K11/18
Sound-focusing or directing, e.g. scanning	G10K11/26
Using reflection, e.g. parabolic reflector	G10K11/28
ELECTRICITY	H
BASIC ELECTRIC ELEMENTS	H01
AERIALS	H01Q
Devices for reflection, refraction, diffraction, or polarisation of waves radiated from an aerial	H01Q15
Reflecting surfaces	H01Q15/14
Curved in two dimensions, e.g. paraboloidal	H01Q15/16
Collapsible reflectors	H01Q15/16B
Inflatable	H01Q15/16B2

A look at the coloured text reveals commonality in the different technical fields

HUMAN NECESSITIES	A
AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY; HUNTING; TRAPPING; FISHING	A01
ANIMAL HUSBANDRY; CARE OF BIRDS, FISHES, INSECTS; FISHING…	A01K
Culture of fish, mussels, crayfish, lobsters, sponges, pearls or the like	A01K61
Floating fish-farms	A01K61/00F
PERFORMING OPERATIONS; TRANSPORTING	B
PHYSICAL OR CHEMICAL PROCESSES OR APPARATUS IN GENERAL	B01
SEPARATION	B01D
Processes of separation using semi-permeable membranes…	B01D61
Reverse osmosis; Hyperfiltration	B01D61/02
Energy recovery	B01D61/06
CHEMISTRY; METALLURGY	C
TREATMENT OF WATER, WASTE WATER, SEWAGE, OR SLUDGE	C02
Treatment of water, waste water, or sewage	C02F1
By heating	C02F1/02
By distillation or evaporation	C02F1/04
Using solar energy	C02F1/14
MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING ENGINES OR PUMPS	F
HEATING; RANGES; VENTILATING	F24
PRODUCING OR USE OF HEAT NOT OTHERWISE PROVIDED FOR	F24J
Use of solar heat, e.g. solar heat collectors	F24J2
Solar heat collectors having working fluid conveyed through collector	F24J2/04
Having concentrating elements	F24J2/06
Having reflectors as concentrating elements	F24J2/10
Parabolic	F24J2/12
Flexible	F24J2/12C

MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING ENGINES OR PUMPS	F
HEATING; RANGES; VENTILATING	F24
PRODUCING OR USE OF HEAT NOT OTHERWISE PROVIDED FOR	F24J
Use of solar heat, e.g. solar heat collectors	F24J2
Component parts, details or accessories of solar heat collectors	F24J2/46
Arrangement of mountings or supports	F24J2/52
Airborne solar collectors, e.g. using inflated structures	F24J2/52D
PHYSICS	G
MUSICAL INSTRUMENTS; ACOUSTICS	G10
SOUND-PRODUCING	G10K
Methods or devices for transmitting, conducting or directing sound in general…	G10K11
Methods or devices for transmitting, conducting, or directing sound	G10K11/18
Sound-focusing or directing, e.g. scanning	G10K11/26
Using reflection, e.g. parabolic reflector	G10K11/28
ELECTRICITY	H
BASIC ELECTRIC ELEMENTS	H01
AERIALS	H01Q
Devices for reflection, refraction, diffraction, or polarisation of waves radiated from an aerial	H01Q15
Reflecting surfaces	H01Q15/14
Curved in two dimensions, e.g. paraboloidal	H01Q15/16
Collapsible reflectors	H01Q15/16B
Inflatable	H01Q15/16B2

1.1.2 Groups of like entities: Multiple Aspect Classification (MAC)

When patents existed uniquely on paper, the possibility of combining different collections to identify a group of documents classified according to multiple criteria cannot have been easy.

Putting the information in computers has facilitated search through the use of Boolean operators. The logical AND can be used to find documents that intersect multiple classifications, the logical OR to add collections together, and the NOT to exclude a collection.

Searching a paperless root-based structure with Boolean operators has been a great improvement.

However computer storage removes the constraint of a root structure.

Families of classes can be used to define a concept in a manner analogous to defining a point in multiple dimensions.

A good example is the F-terms designed by the Japanese Patent Office to operate with, and supplement, the IPC [6].

The example used in Wikipedia is reproduced below. The four 'viewpoints' as they are called, relate to the ingredients, cuisine, cooking and special classes.

9Z999 Dishes (X99Y 1/00—1/12)

AA INGREDIENTS
AA11 Meat
AA12 . Beef
AA13 . Pork
AA14 . Lamb
AA21 Seafood
AA22 . Fish
AA31 Vegetable

BB CUISINE
BB41 Asian
BB42 . Chinese
BB43 . . Cantonese
BB44 . Turkey
BB51 European
BB52 . French
BB53 . Italian

CC COOKING
CC11 Boiled
CC21 Fried
CC31 Roasted
CC41 Steamed

DD SPECIAL
DD11 Halaal

DD12 Vegetarian

And so a dish can be classified as Fish, French and Steamed; or as Beef, French and Fried.

Such a system provides a great number of classification options. The viewpoints can be thought of as dimensions, which can pinpoint information in a manner analogous to triangulation in navigation.

Such viewpoints can also operate locally on a root of the IPC. Alternatively such a scheme could operate independently of a root structure, and in theory could operate across the whole of technology.

The B82Y codes of the IPC developed for Nanotechnology, and the Y02N codes developed by the European Patent Office for Climate change mitigation technologies, do the latter [7] [8]. That said they are designed to support the IPC, and are not an alternative to it.

Durham Zoo will refer to a system of multiple families of classification codes operating globally across technology as Multiple Aspect Classification (MAC).

1.2 The language of patents and the language of academia and industry

A computer disk drive is not called a disk drive in the IPC.

In the IPC, disk drives are classified in ‘information storage based on relative movement between record carrier and transducer’.

The IPC group includes all manner of ‘moving’ memories such as punched tape, magnetic tape, nano-cantilever and vibrating-around-a-central-point.

An invention in ‘moving’ memories may be of broader application than a disk drive. Non-specific terms are thus routinely used in patents so as not to restrict the scope of protection. A disk drive may be referred to simply as a ‘storage device’.

Thus whilst the IPC does a good job in grouping together similar things, the esoteric language used may not be readily understood by the non-patent practitioner.
A search of the IPC classification for ‘disk drive’ with ‘natural language processing’ returns a zero result [9].

A search of the IPC classification for ‘disc’ returns 30 results, which includes entries for disk drives but many things besides. It is not easy to make sense of the results, even for an experienced patent practitioner.

Durham Zoo however guides a classifier with a specific example of a concept, described in the language of industry, to a zootag that may define the concept in more broadly applicable terms, this in a manner analogous to the IPC.

Just a tip to make sense of patent speak: a patent may be more easily understood from an academic publication of the same subject matter.

1.3 Code-mixing

Classification and database structure need be compatible. This is more likely to be the case if both are designed at the same time.

MAC-style combinations of codes used to define multiple inventive concepts in a single document should have independent entries in the database, this to reduce noise.

The necessity can be demonstrated by considering a simple example.

If a document has 3 different concepts, each of which is described by a combination of 2 codes, there will be 6 classes attributed in total to the document.

If the concept code pairs are not separated, the document would be retrieved by all-and-any of the fifteen combinations of the classification code pairs, of which twelve will be false combinations.

Put another way, if the three concepts exist as code pairs A with B, C with D and E with F, it would be advantageous if the codes were separated to avoid the document being retrieved erroneously in a search for A with C.

A ridiculous example may be derived from the fish-farm patent. Codes for fish and farm, and inflatable and reflector would return the patent in a search for a fish reflector.

A potentially noisier, and more insidious example would be the creation of codes defining parts of a bicycle: aluminium and frame, and carbon and wheels?

D Z separates the different zootag combinations into different entries that will not interfere.

1.4 Future patent office schemes

The IPC, designed as a global standard and translated into many languages, was not implemented as such across the major patent offices. This has resulted in incompatibility between schemes.

The 'fiveIPoffices', five major patent offices from around the world, are seeking to harmonize their classification in a common hybrid classification (CHC) [10] [11].

The European Patent Office and the United States Patent and Trademark Office have decided to implement a common patent classification (CPC) scheme in order to promote a CHC [12].

The integration of classification information developed by the Japanese Patent Office to the CPC is also foreseen [13].

We believe that the Durham Zoo design is inherently more powerful than the proposed schemes.

That the IPC and its derivatives are operational and are the fruit of a massive human investment is acknowledged. To ‘start again’ with a classification scheme and to reclassify the documents that are already classified would be foolish.

However Durham Zoo can incorporate all and any classification information completed to date in an elegant and relatively inexpensive way. Even very different schemes can be combined together on ZSD’s. The solution to the problem is provided later in section 5.6.

1.5 More shades-of-grey

We fully understand that the F-term viewpoint example from Wikipedia was not intended for rigorous academic scrutiny, it is a good example to demonstrate a concept, however a closer look can demonstrate the limitations of ‘black- and-white’ classification in a MAC context.

For example, Turkish cuisine appears under Asian cuisine.

But Turkey, whilst in large-part in Asia, is also part in Europe. According to Wikipedia, Turkish cuisine is a ‘fusion and refinement of Central Asian, Middle Eastern and Balkan cuisines’. Whilst this cannot be described in MAC it can be represented on a ZSD.

And how would English cuisine be represented with MAC, which whilst having evolved greatly in recent years still has the classics of fish and chips, roast beef and Yorkshire pudding, and chicken tikka masala?

According to Jamie Oliver, and he would know, fish and chips is Portuguese Jewish in origin. Did Spinoza ponder his philosophy over a bag of fish and chips?

More and more codes can be created in a MAC to make finer and finer divisions within a family, however there is nothing in MAC to represent any inference or association between the different entries.

2. What and how to classify

A working system will require the development of tutorials for classifiers, with worked examples from different technical fields. The following explanation introduces the main concepts.

2.1 The source

There are different types of written technical disclosure: there are patent applications and granted patents, there are journal articles, conference proceedings, there is company literature including white papers, Bachelor's degree projects, PhD thesis, web forums and books.

Good ideas can come from surprising sources: a Donald Duck cartoon strip was used as an anticipation for a patent application for raising a sunken ship with ping-pong balls [14].

From a patenting perspective the 'prior art' is not restricted to the written form: anything that has been made public in any language and by any means is included. The spoken word may be particularly pertinent with regard to 'indigenous' or 'traditional' knowledge [15].

It is the intention to be able to classify anything that is part of the prior art and that can be referenced. It may be useful to provide categories for the different types of disclosure.

2.2 The content

The subject matter of individual disclosures can be of different kinds. Many disclosures introduce something 'new'. It is this 'distinguishing content' that needs to be classified.

For recent disclosures the above requirement is absolute. For old or ancient disclosures, the distinguishing information will be that which was distinguishing at the date of publication.

There may be more than one disclosure which has broadly the same distinguishing content and which is deserving of classification.

However it is not the goal to classify a same and well-known concept to the extent that the original disclosures are lost in the mass.

If a same concept is further refined it is hoped that the library of zootags is able to define the new and distinguishing content. If a large collection of documents is described with the same zootags, it may well be that the library of zootags needs to be expanded and enhanced.

Further as regards distinguishing content: a patent application may eventually prove not to disclose something new, but a granted patent definitely has newness as a formal requirement. Furthermore, what is new in a patent application also needs also to be ‘inventive’ to be deserving of a grant.

Inventiveness is that which is non-obvious, alternative or clever about what is new. Inventiveness requires a comparison with what exists in the prior art and the judgement of a ‘skilled person’.

In Durham Zoo we believe it both reasonable and pragmatic to *a priori* accept the appreciation of the author in good faith.

The classifier needs to find the best fit of the distinguishing content with the available zootags. It is important that the terminology of Durham Zoo is implemented and not the terminology of the disclosure.

In most cases the distinguishing content as identified by the author will be the same as that identified by the classifier (and this will be so in all cases where the author *is* the classifier), however a classifier may identify and classify additional distinguishing information if they deem it worthwhile.

There may be cases where there is no distinguishing content. A review of known techniques or the state of the art in a specific field may be a very useful and informative disclosure *per se*, but may be of little interest in terms of distinguishing information. To classify the reviewed content would be duplicitous, potentially onerous, and may create noise.

However where a classifier believes part of the content to be of particular interest such information can be classified. It may also be that a comparison of the reviewed content is of interest. And it could be that review documents are given a specific category identifying them as such.

2.3 The degree of specificity

Classifiers will also be required to judge to what level of *specificity* a concept should be classified. A concept may be described in the context of a single *application*, but it may be more broadly applicable.

For example, a disclosure relating to a saddle for a horse may be of more general, or more specific use.

A saddle adapted for a particular horse or particular use, for example for carrying ceremonial drums on a dray horse, would likely be more specific than a general-purpose saddle.

If the distinguishing information relates to the distinctness of the design for the specific application, then the distinctness should be represented in the zootags. An application zootag ‘saddle for ceremonial drums on a dray horse’ may be then be appropriate.

However it may be the case that a concept described in the context of a specific saddle is of more general application. A ceremonial dray horse saddle manufacturer may disclose distinguishing content relating to a particular material that would be applicable to all saddles equally as well.

The correct level of application zootag may depend on what was explicitly disclosed as regards the extent of the application. If the author describes the material as being used for a dray horse saddle, but that the applicability would extend to all saddles, it would be preferential to zootag with a generic saddle.

If the author has described the material as being used for a dray horse saddle but the extension of the use of such a material to a generic saddle *is not obvious*, then the application should remain limited to a dray horse.

And depending on the zootags, application to a ‘generic’ saddle could mean for horses, and most likely to donkeys, but what about to camels, or elephants? And what of saddles on bicycles? A material may find application to all saddles, and maybe more besides.

The extent of the applicability of the concept requires the judgement of the saddle expert. What exactly has been disclosed, and what is implicit? The saddle expert then has to find the best fit with the existing zootags.

The consequences of attributing a zootag of a more general or a more specific nature are discussed below. For simplicity we have decided to define zootags for a ‘generic’ saddle, an equine saddle, and a bike saddle.

A search for any saddle will return all the saddles with the full weighting in the application dimension given that all saddles are examples of the generic saddle. A search for a particular saddle will rank the different saddles.

Whether a particular document is seen during a search or not depends on its’ ranking in the results and how many documents are reviewed before the search is cut off.

The ranking may of course be dependent on other dimensions and so a more general or more specific classification of the application dimension may be of little practical consequence.

However if the distinguishing information of a disclosure is classified as a generic saddle when it actually relates to an equine saddle, the equine nature of the disclosure is lost.

During a search for an equine saddle our document risks being ranked below any other equine saddle zootagged documents. The risk is of losing, or more exactly of not seeing the document.

If the distinguishing information of a disclosure is classified as an equine saddle when it actually relates to a generic saddle there are other risks. The search of the distinguishing information for a *different specific* saddle, for example the bike saddle, risks the equine saddle zootagged document being ranked too lowly.

Thus there exist risks either way. The classifier must use their expertise to judge the correct level of applicability. On balance, the unduly specific classification would appear to be the lesser of two evils.

The inherent power of the multidimensional search may eliminate much imprecision.

Regarding the specificity of patents: in order to confer protection in as broad a manner as possible a patent claim may be worded in a very unspecific manner. The wording may be different from that routinely used by practitioners in industry.

As an example, Professor Morse’s telegraph patent ‘for making or printing intelligible characters signs or letters at any distance’ was ‘claimed’ in too broad a manner. The disclosure explained one way of transmitting a message, but not all the possible ways of doing it. It was subsequently amended.

To *classify* the telegraph concept as drafted in the original broad claim would not be a good idea: the risk would be that the Morse patent would be ranked too lowly in a search for the telegraph concept.

2.4 More than one concept in a document

This is another important point. As was seen in the fish farm patent different classification information was added to a single document.

This may be for a same technique applied to different applications: for example G01K11/28 relates to a parabolic reflector for directing sound, and H01Q15/16B2 relates to a parabolic reflector for directing waves.

Thus for the fish farm patent to have multiple and different zootags for the application dimension would be perfectly correct.

However it may be that there is very different ‘distinguishing information’ in a single document. Patent applications commonly bundle together information that cannot be construed as having a common concept. In patent terminology such documents may be ‘non-unitary’.

In such a case it would be quite correct for a single document to be zootagged with multiple zootag combinations that have little or no overlap.

It may be a good idea to ‘cluster’ together different zootags relating to a same concept. Again a database field including an explanatory text as to where in the document and why the particular zootags have been added would appear a good idea.

3. More about the fuzzy search engine

3.1 The search engine calculations

The weighting of the related entities on the ZSD is a function of their position.

When seen on the horse ZSD, the donkey zootag appears as a unique entry. However the donkey zootag has its own ZSD. There are different ways in which the donkey zootag entry on the horse ZSD can be interpreted during a search with the horse zootag.

The simplest way is to include everything that is fully a donkey, thus anything with the donkey zootag, and anything zootagged with any of the specific Examples of donkeys, the Examples appearing underneath the (generic) donkey zootag on the donkey ZSD.

A more complex method entails using all the entities on the donkey ZSD: i.e. those above and below the donkey zootag. In this way the horse ZSD would in fact address the ZSD's of all the entities it refers to. The ZSD refers to a plurality of ZSD's, which in turn refer to each of their ZSD's. It can become a bit 'fractal' like.

Before discussing the details further we need remind ourselves of the notation for the elements on a ZSD.

The horse zootag on the horse ZSD is the 'ZSD Subject'.

Those entities below the ZSD subject are all 'Examples' of the ZSD subject. Examples are not inferred, given that they are Examples of, and not approximations to, the ZSD Subject.

Examples may appear on a ZSD in different levels of subdivision, however this does not affect their weighting as regards the search engine: they are all the subject entity to the full extent.

Those entities above the ZSD subject are inferred entities and called 'Inferrands' in Durham Zoo parlance.

Inferrands are inferences by degree; the number of degrees dependent on how many individual inferences need be made in order to get from the ZSD subject to the Inferrand in question.

Those entities linked directly to the ZSD subject are called first-degree Inferrands. Those Inferrands that are inferred from the first-degree Inferrands are called second-degree Inferrands, and so on…

The figure below shows the examples and the first-degree Inferrands. The ZSD hierarchy and second-degree Inferrands will be apparent from the example that follows.

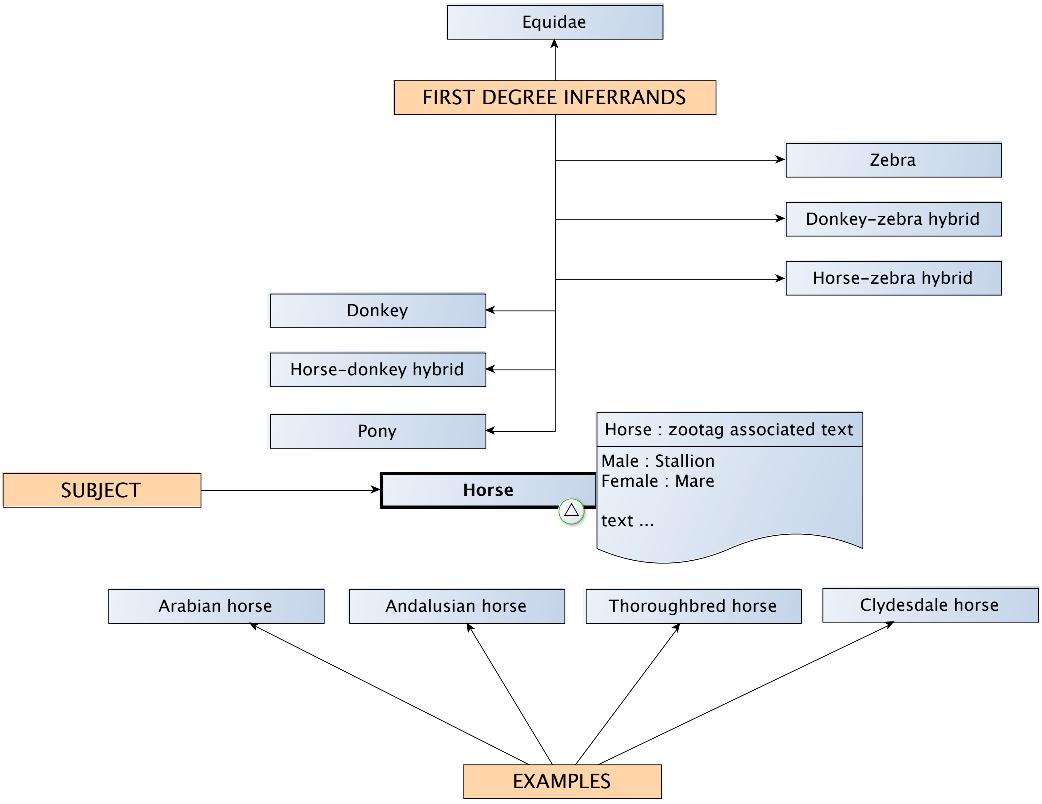

Fig. 5 Horse ZSD with examples and first-degree Inferrands

3.1.1 The simplex search

The simplex search includes the ZSD Subject and all Examples of the ZSD Subject.

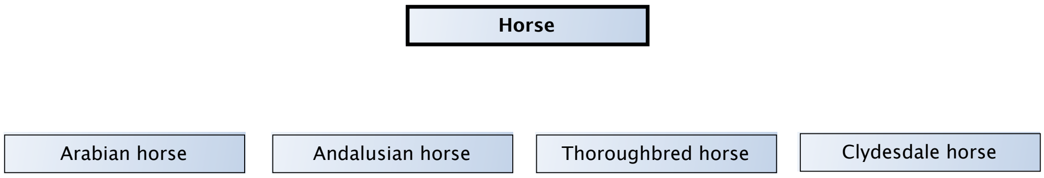

Fig. 6 Simplex search

3.1.2 The complex search

In a single complexity search the first-degree Inferrands and the Examples of the first-degree Inferrands are additionally included.

The weighting value of each first-degree Inferrand and all its associated Examples is the same, being a function of the position of the first-degree Inferrand on the ZSD.

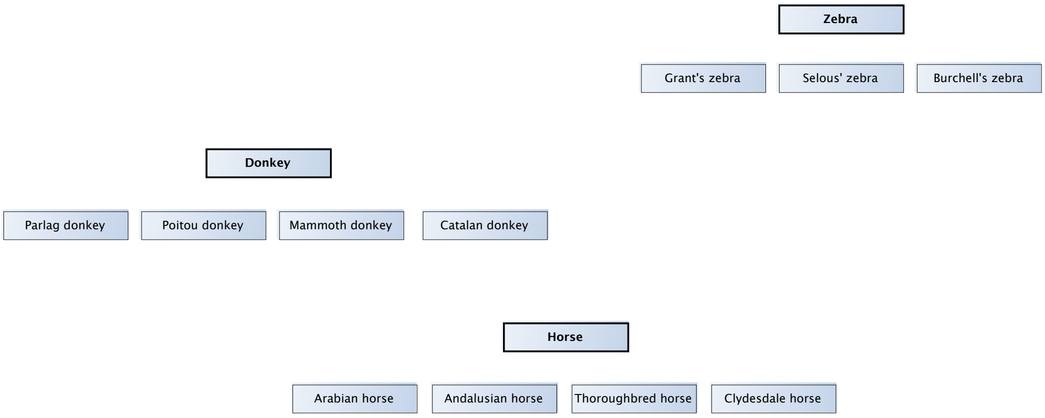

Fig. 7 Complex search: single complexity

In a double complexity search the second-degree Inferrands and the Examples of the second-degree Inferrands are additionally included.

The weighting value of each second-degree Inferrand and all its associated Examples is the same, being a multiplication of the weighting value of the first-degree

Inferrand on the ZSD with respect to the ZSD Subject, and the weighting value of the second-degree Inferrand with respect to the first-degree Inferrand.

In the above manner the search engine simply uses the ZSD's of the Inferrands in a 'fractal' like manner: additional levels of complexity simply 'zoom in' and import the inferences from the addressed ZSD's.

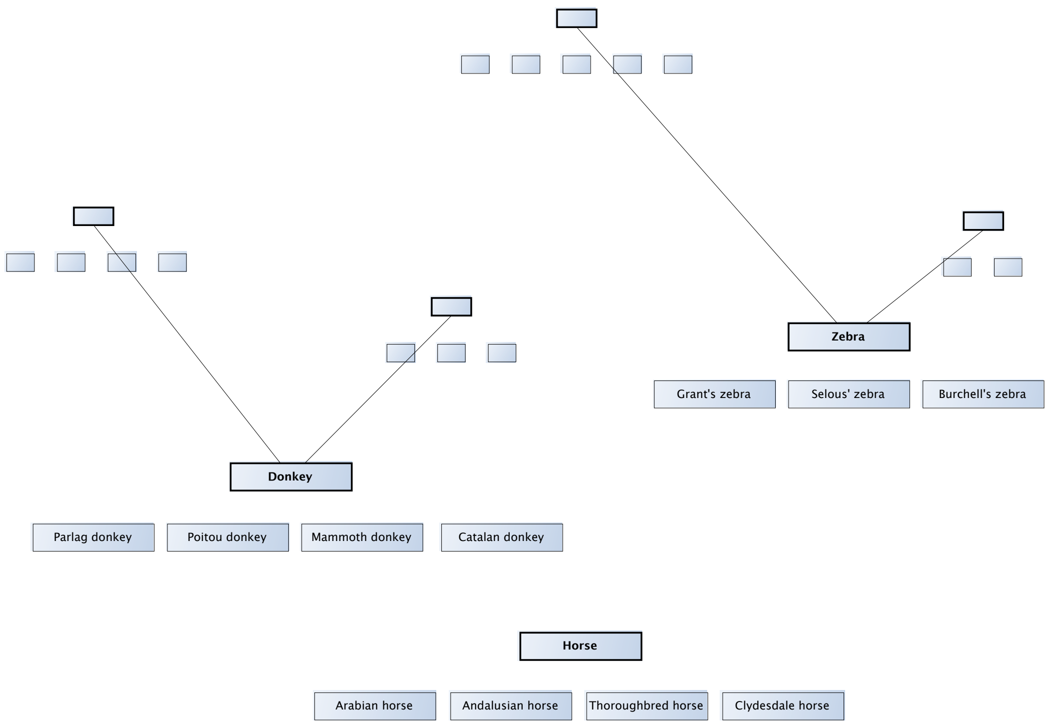

Fig. 8 Complex search: double complexity

The complexity could be extended further, maybe down to a minimum threshold. However the utility of too many extensions may be of little value in the context of a search in five dimensions.

3.2 Duplicates

Complex searching can result in seeing a same zootag multiple times in the calculation process. Although it would be possible to calculate an inference from the multiple appearances, we believe a better method is to simply decide on which one to take.
There are two types of duplicates: special-case Examples and true duplicates.

3.2.1 Special-case Examples

It could be that there is a particular species of donkey that resembles a horse to a much greater degree than the 'generic' donkey. This particular species of 'horse-like donkey' is thus poorly defined as a 'normal' Example of a donkey on the horse ZSD.

The 'horse-like donkey' zootag should thus receive special treatment and be placed lower in the horse ZSD than the generic donkey zootag to reflect its greater similarity to a horse.

This would be a 'special-case' Example of a donkey given that the weight on the horse ZSD would be something other than the generic donkey with its more standard Examples.

A complex search would thus see the appearance of the horse-like donkey zootag as a first-degree Inferrand on the horse ZSD, and then again as an Example of the generic donkey on the donkey ZSD.

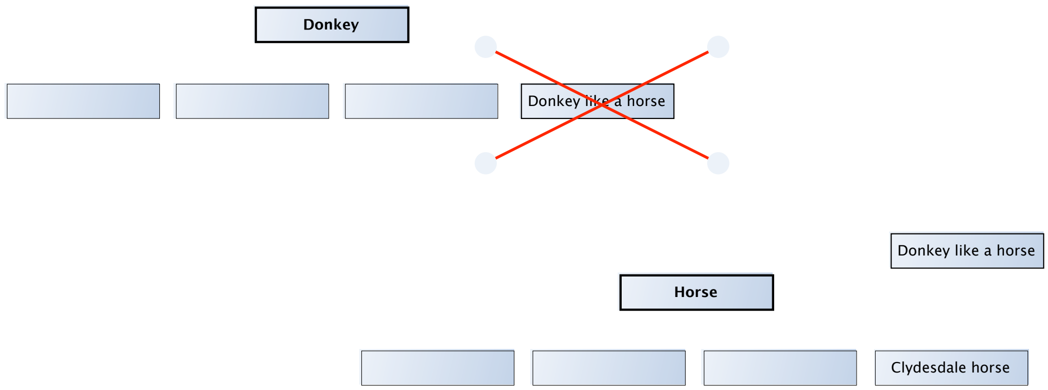

Fig. 9 Duplicate removal 1

The information we need is the inference as seen directly from the horse.

Thus the horse-like donkey Example underneath the generic donkey zootag needs to be filtered out.

Similarly, a 'horse-unlike donkey' would appear above the generic donkey zootag. Again the first-degree Inferrand is the correct one.

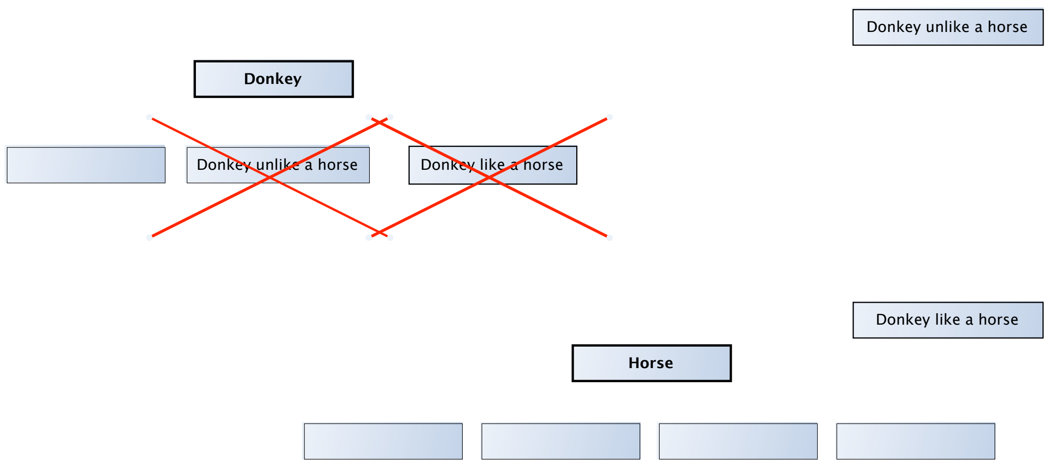

Fig. 10 Duplicate removal 2

(We understand that with this methodology the simplex search is an approximation as regards the handling of 'special case' Examples, however we believe the approximation is valid for a first approximation.)

3.2.2 True duplicates

It could be that a same zootag appears via a different sequence of inferences.

Again on our horse ZSD, a hybrid between a horse and a donkey (‘horse donkey hybrid’) could appear as an Example of a first-degree Inferrand ‘horse hybrid’ zootag, and it could also feature as a second-level Inferrand via the ‘donkey’ zootag.

It could be argued that an entity should have a single inference value with respect to any other entity. However the horse and donkey ZSD’s are developed from different perspectives.

The lowest degree Inferrand is taken as the best option using an Occam’s razor-like assumption.

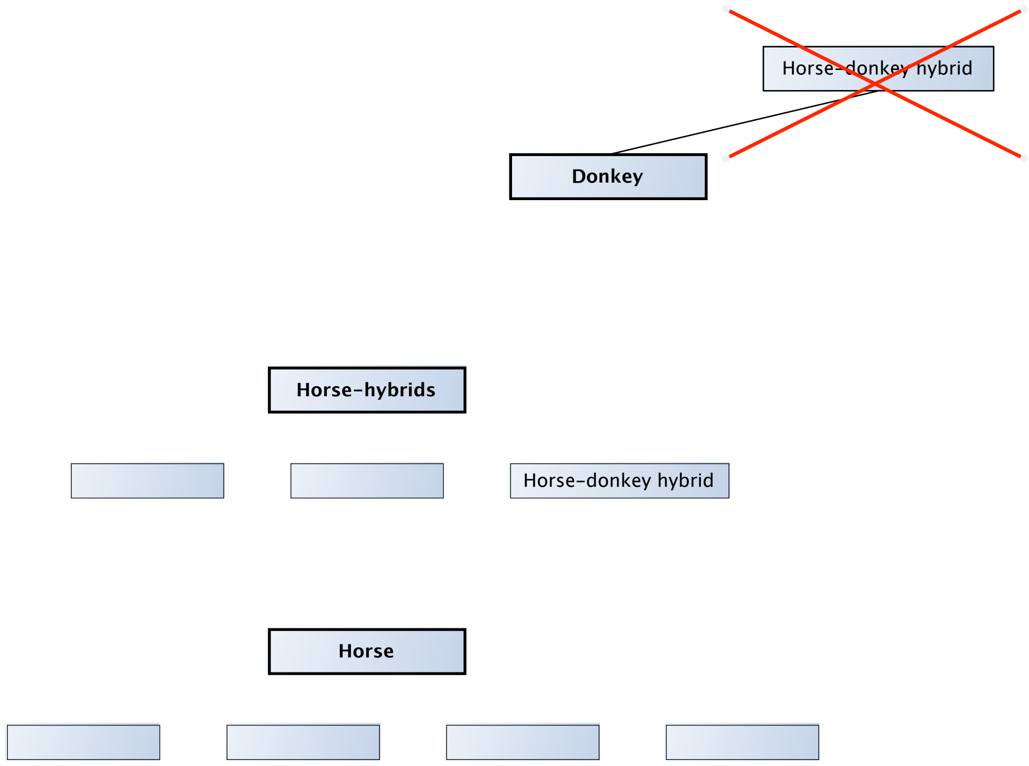

Fig. 11 Duplicate removal 3

And wherever the same entity appears with the same degree, the higher weighting is the one selected. This would appear justified, given that there is more than one independent inference.

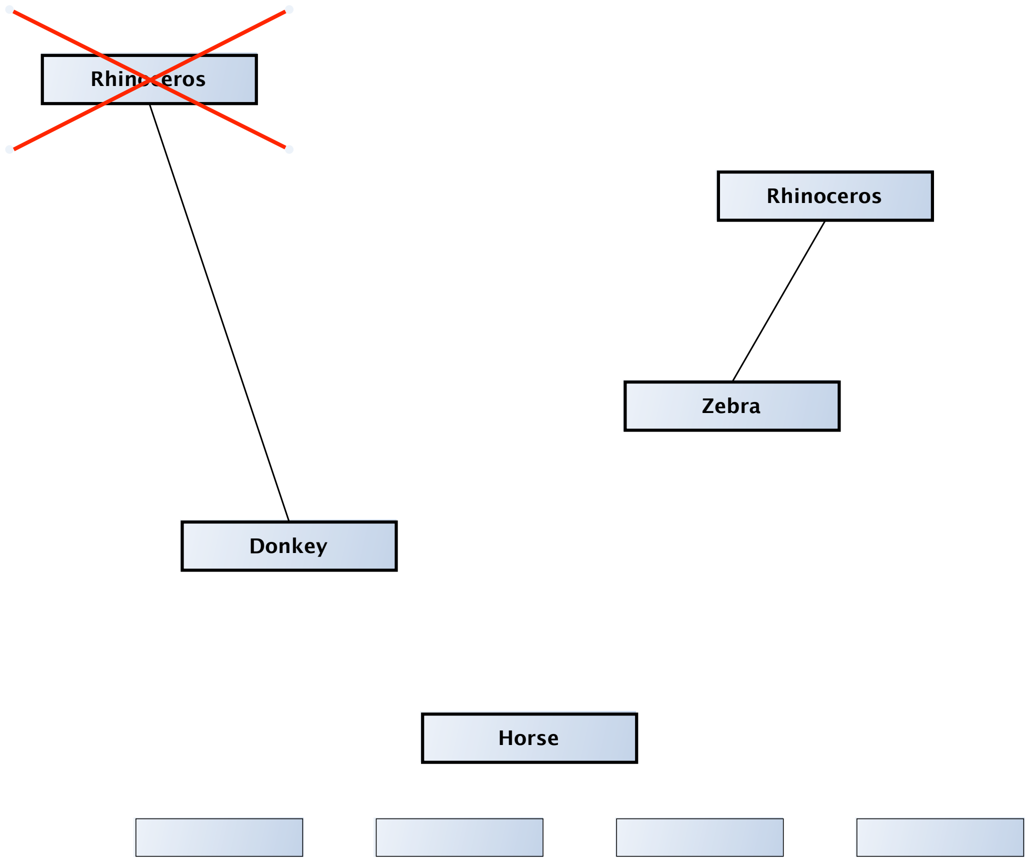

Fig. 12 Duplicate removal 4

3.2.3 Complete duplicate removal procedure

Thus the complete procedure is as follows:

a) simplex search:

i) read the Subject zootag from the ZSD
ii) read the Examples of the subject zootag
iii) assign the full weighting to all

b) single complexity extension 1:

i) read the first degree Inferrands
ii) where a first degree Inferrand is a duplicate from a) ii), replace the full weighting with the inferred weighting (remember, this procedure allows an atypical horse to be less of a horse than a generic horse)

c) single complexity extension 2:

i) read the Examples of the first degree Inferrands
ii) where an Example of the first degree Inferrand is a duplicate from a) ii), replace the full weighting with the inferred weighting
iii) where duplicates appear from c) i) take the highest weighting and reject the other

d) double complexity extension 1:

i) read the second degree Inferrands
ii) where a second degree Inferrand is a duplicate from a) ii), keep the full weighting using the Occam's razor-like assumption
iii) where a second degree Inferrand is a duplicate from c) i), keep the first degree Inferrand weighting using the Occam's razor-like assumption.
iv) where duplicates appear from d) i) take the highest weighting and reject the other

e) double complexity extension 2:

i) read the Examples of the second degree Inferrands
ii) where an Example of a second degree Inferrand is a duplicate from a) ii), keep the full weighting using the Occam's razor-like assumption
iii) where an Example of a second degree Inferrand is a duplicate from b) i), keep the weighting of the first degree Inferrand using the Occam's razor-like assumption
iv) where an Example of a second degree Inferrand is a duplicate from c) i), keep the first degree Inferrand weighting using the Occam's razor-like assumption.
v) where duplicates appear from e) i) take the highest weighting and reject the other

Thus where duplicates appear, the search engine will make an intelligent choice.

The handling of duplicates in Durham Zoo also facilitates the development of the classification scheme. Different ZSD's can be developed independently and simultaneously: the search engine will handle any repeat entries.

A ZSD can address other ZSD's and the information in them as a first approximation. As special-case Examples or true duplicates become apparent, information can be added and further refined by the promotion of subject Examples, and second-degree Inferrands and/or their Examples, to first-degree Inferrands.

Even when the 'correct' information from an inferred ZSD has found a place on the mother ZSD there are no coherency problems with the further development of either ZSD.

That said zootag deletion would need to be handled in accordance with the protocols for zootag evolution mentioned later in section 5.

3.3 Search engine advanced

In its standard configuration the search engine will rank hits as calculated by the five dimensions.

However the ability to fine tune a search is provided by additional functionality at the search stage, this using enhanced features of the search engine.

The order in which a retrieved group of documents are then displayed should be configurable according to user-defined preferences. This is the display engine.

Similar functionalities could be included in both search and display engines.

3.3.1 Searching with fewer dimensions

A search could be restricted to any number of dimensions. And it should be possible to add or remove dimensions in the refinement of an initial search.

However, how are we to compare different documents, hit in different numbers of dimensions? How to compare a near hit in two dimensions with a less accurate hit in three?

The inclusion of more or fewer dimensions has consequences for the type of documents that will be retrieved.
For example a search including dimensions for a problem and a solution may rank a document from a different technological field much higher than a document from the same technological field. It may be that a solution to a problem in an electrical motor is the solution to the same problem in a computer.

The intention may be to find the state of the art in a particular field, or to search prior art from elsewhere.

A searcher wanting to improve the ranking of documents from the field of computers would simply need to introduce the most general code for computers as the application dimension into the search query. This would effectively give a full 1 weighting in the application dimension effect to all documents from the computer field.

Doing so would effectively decrease the pertinence of the motor document in the result without excluding it. However if the problem and solution were unknown to computing the motor document may still be ranked highly.

3.3.2 Variable weighting of dimensions

Until now the different dimensions have all had the same weighting. However this need not be the case.

By varying the weighting of individual dimensions it is possible to put emphasis on one or more of the dimensions in the search query.

In the computer and motor example above, a more radical means for excluding the motor document would be to increase the weighting of the computer application code.

If the weighting attributed to the computer application code were given a value of 1000 instead of a 1, documents from outside the computing field would not get a look-in. The 1000 weighting for the computer application code would effectively translate as ‘must have a computer’.

In a similar manner an effect similar to the Boolean logic NOT operator, the ‘must not have’, could be assigned with a ‘negative’ inference level for that dimension.

Thus in a search for alternatives to a known solution to a particular problem, those documents with the known solution could have a -1 in the solution dimension.

Again with variable weighting this could be magnified to -1000 to ‘harden’ the preference.
The examples above use distance from Euclidian geometry and two alternatives in relation to the weighting: the plus/minus and the stretch.

There may be other and better alternatives? Alternative inner product spaces and numerical analysis techniques could be used to improve search and speed up the calculations.

3.3.3 Crossover between dimensions

As proposed so far, links can only exist within a dimension; however ‘inference’ could be extended across the different dimensions.

For example, the problem of ‘radiation-induced errors’ is a known problem in Dynamic Random Access Memory (DRAM) technology. There will be many documents that have both aspects.

This relation could be used to effectively link the two dimensions. It could be that a search on the problem of ‘radiation-induced errors’ include automatically, or maybe suggest, a weighting factor to literature pertaining to a DRAM application.

Inferences could be suggested or implemented automatically.

It may be helpful to perform a search in one dimension, and then perform a statistical analysis of other dimensions within that group.

In a similar manner there could be dimension crossover between the problem and solution dimensions. Turning a problem into a solution is a well- established method for making good use of an observed behaviour.

The Atomic Force Microscope suffered the problem of atoms being lifted from the surface of the material being investigated onto the microscope tip. This problem was later the solution to the selective removal of atoms from a surface [16].

Another example was the non-permanent stick of the post-it sticker [17].

A problem is often related to a particular situation. It may be that a solution to a very big problem has a smaller problem associated with it. The newer and smaller problem needs to be defined as closely as possible in its new context, and that could include the solution to the original bigger problem.

Any design will have to consider the need for accuracy in the zootags and the need for commonality of the zootags across technology. Further consideration will be the appropriate complexity for the crowd and the cost in terms of number crunching.

3.3.4 Multiple codes in a single dimension

Up to now we have described a concept using a single code in each of the MAC dimensions (technology, application, operating mode, problem and solution). However the Durham Zoo search engine should accept multiple codes in any of the dimensions.

However the search engine needs to know how to interpret such inputs.

The simplest case is to attribute a logical OR to the different codes. This is of use where any one of multiple possibilities needs to be taken into consideration.

However it could be that multiple codes in the same dimension are used to improve the specificity of the query. Technology code A *with* technology code B would be a logical AND.

The AND is more problematic for the search engine calculations. Would the results of hits with code A and code B be simply summated?

Such a procedure would lead to a potential hit of magnitude 2, which would effectively increase the weighting of the dimension in question, assuming that the other dimensions were searched with a single entry.

This may be considered entirely logical and entirely reasonable: if more information is hit in a dimension it follows that this increase in specificity is reflected in the ranking calculations.

However this reasoning could stretch too far and unduly distort the results, especially if a third or a fourth input were included.

Hits with multiple codes could be summated and re-weighted to a maximum of magnitude 1. However this would require the re-weighting of single code hits. Also the ‘added value’ of multiple code hits may be deemed as being under-valued in such a scheme.

A compromise solution would be variable weighting within a dimension. A first code hit would be weighted as 1, a second code hit as 0.4 and a third hit as 0.2. In logic terms this could be referred to as a ‘variable weighting m-out-of-n’.

Similar consideration needs be given to the OR’ing of multiple codes.

The incorporation of a ‘dependency graph’ linking different zootags is another area that should be investigated, especially in relation to the problem dimension [18].

Another option would be to define single dimensions in a manner analogous to a concept. For example the problem dimension could itself be defined as a concept with n sub-dimensions. Again the problem dimension is a strong candidate for this approach.

3.3.5 Bespoke but compatible

We believe it is likely that different technical fields will be best served by adaptations to the basic scheme.

Just as ‘applications’ are written for the Apple App Store, so code designers could design ‘classification apps’ for specific technical fields. A ‘Durham Zoo development kit’ could be considered.

Any design should provide increased functionality but not at the expense of commonality in search across technology.

It may also be possible for searchers to edit ZSD’s according to their particular needs prior to using them in a search.

3.3.5.1 Multiple inputs in ‘pods’

Many entities, such as the composition of a product, require multiple descriptors. A pod is a proposed shorthand structure for grouping multiple descriptors.

Standard notation will likely need to be developed to serve specific fields. That said the design should be mindful of the benefits of maintaining commonality across the database and the ZSD structure.

By way of example, consider the structure of laminates:

An individual layer may be described by its composition and thickness, such as material x + material y + material z, with a thickness in n mm. This could be represented in the following notation.

(x, y, z, n)

The notation could be extended to include percentages of the composition.

Maybe a pod could be developed to describe a necessary Boolean construct, such as (A AND B) OR (C AND D). For now it remains an open question.

3.3.5.2 Sequence search

In many concepts it is the order in which particular steps are undertaken that define the particularity of an idea. The individual steps may be unremarkable when considered individually.

This is the case in a range of applications from chemical processes, to database recovery methods, to the laminate structures mentioned earlier.

Below is an idea for defining a sequence.

[A, B] means A must come before B

[[A, B]] means a must come directly before B, i.e. with nothing in between; hereafter referred to as indivisible.

As an example of sequence definition with these two operators:

[1, 2] 3, 4, 5 = 1 before 2; with 3, 4 and 5 at any time

[[1, 2, 3]] 4, 5 = 1 before 2 before 3 indivisible, with 4 and 5 at any other time

[[1, 2, 3]] [4, 5] = 1 before 2 before 3 indivisible, and then 4 before 5

As a further example the following sequences 1-30 include 3 and 4 variables:

1. ABC
2. ACB
3. BAC
4. BCA
5. CAB
6. CBA
7. ABCD
8. ABDC
9. ACBD
10. ACDB
11. ADBC
12. ADCB
13. BACD
14. BADC
15. BCDA
16. BCAD
17. BDAC
18. BDCA
19. CABD
20. CADB
21. CBAD
22. CBDA

23. CDBA
24. CDAB
25. DABC
26. DACB
27. DBCA
28. DBAC
29. DCAB
30. DCBA

Below are listed four examples of those sequences that correspond to a searched sequence:

i) A, B, C = 1-30

ii) [A, B] C = 1, 2, 5, 7-12, 19, 20, 24, 25, 29

iii) [A, B] [C] = [A, B, C] = 1, 7, 8, 11, 25

iv) [[A, B]] C = 1, 5, 7, 8, 19, 24, 25, 29

The NOT operator should again be included in a working notation.

3.3.5.3 Pods and sequences together

The combination of pods and sequences should be included in the functionality of Durham Zoo.

However again such an input may result in a large amount of number crunching.

It may be necessary to perform such a complex search in an iterative manner where increased precision is searched in a progressively restricted number of documents.

That said, the calculations would be well served by parallelism, which could be implemented in a parallel-processor hardware architecture.

4. Durham Zoo protocol

4.1 Zootag scheme development

Before classifiers can classify in Durham Zoo, the zootagging schemes need to be developed. And as stated earlier the schemes will always be evolving, always a work in progress.

The following are a few ideas about how to run the scheme.

To design and update a classification scheme requires a deeper knowledge and level of expertise than simply using one. To design a zootag that is consistent with other zootags, that is readily understandable, and able to accurately and unambiguously define a concept requires a considerable intellectual investment.

And as the NASA Survival Scenario Exercises adequately demonstrate a team nearly always makes a better decision than an individual.

Thus the management of each zootag in Durham Zoo will be entrusted to a Codekeeper, aided by a small group of Sages.

A same team would manage all the zootags and ZSD’s in a specific technical field. As the numbers of zootags increases so the team could be enlarged and then sub-divided to manage the increased workload, and to provide increasingly specialised expertise.

Codekeepers and Sages would need to keep abreast of developments in related fields. A supervisory board may need to be established to manage the bigger picture.

Proposals to edit any ZSD will be discussed and eventually voted on by the responsible Codekeeper and Sages.

There should be an even number of Sages and one Codekeeper who will have the casting vote. The Sages should be experienced classifiers, preferably with a range of expertise and experience from academia, industry and the patent world.

Whilst English appears to be the natural default language of Durham Zoo, the Sage’s should also be enlisted taking account of language skills, and especially where pertinent to the technological field in question.

For many technical fields this may include Chinese, Japanese and Korean as well as major European languages such as French and German. Translation of the zootagging scheme into languages other than English will be necessary: the skill level and hands-on involvement of Sages make them ideal candidates for the job.

The operation of Durham Zoo should be as efficient as possible. Thus whilst input regarding revision of the zootagging scheme should be actively encouraged, we would suggest the use of quorum voting and silent procedure should be encouraged wherever possible.

Classification is not an exact science. For different parties to hold different opinions about a matter is both perfectly normal. And a vigorous debate can be very positive.

The goal in matters classification is the adoption of a commonly held point of view. Which particular viewpoint is adopted is commonly of less importance than the consistency of the viewpoint. Thus energies should be focused on finding consensus.

‘Mission statements’ and codes of ‘principles and values’ are mostly common sense. That said referral to them can be useful when there are differences of opinion.

In matters classification: nobody wins an argument but everyone wins in consensus.

4.2 Zootagging I/O

4.2.1 Managing the input
‘Death alone is silent’ (Jacques Attali)

Controls on the zootagging need to be less stringent than those for the development of the ZSD’s. Durham Zoo needs to be open enough to encourage meaningful input. The consequent variation in information of an open system can be managed by a combination of tools that can improve the credibility of information, and tools that can select and prioritise information.

In information theory there is information and there is noise: good classification is information, whereas poor classification is noise.

And just as you can’t make an omelette without breaking eggs you can’t classify without creating noise. So can noise be controlled?

Access to zootagging could be restricted to those competent in the theory of classification, accepting of operating protocols that value consistency above ‘academic rigour’, and having ‘signed up to’ the mission statement and code of principles and values.

Access could be conditional on completing an on-line tutorial, with worked examples. Input to Durham Zoo could then be controlled using e-mail addresses and/or identification and passwords.

Even with the above-mentioned controls it would be naïve to believe that the zootagging data would be totally consistent. More realistic is to think in terms of the quality of the information.

‘Matters of fact’ are regularly contested and differences of appreciation are likely to be more commonplace. Thus even excluding the inevitability of mistakes, it would be impossible for zootagging to represent an absolute truth or be 100% consistent.

But then the same could be said of Wikipedia, which nonetheless remains an incredibly useful source of information that has grown in credibility as it has developed.

True, the accessibility of Wikipedia has made it prone to ‘attack’, and yes it would be naïve to think that a Durham Zoo database would be immune from malicious intentions. But information systems can allow edits to be ‘undone’ or ‘rolled back’ to remove content. And the crowd has many pairs of eyes.

Citizendium bills itself as a more credible alternative to Wikipedia, and there is indeed more control on posts. That said it is much less known.

4.2.2 Managing the output

The ability to manage the output is related to the search engine functionality.
The following information has been included here in the Durham Zoo Protocol as it includes user interaction.

Searchers need be able to *select and prioritise* documents according to who has performed the classification. Searchers need also be able to select and prioritise the documents they review according to the classification information itself.

Preferences may be applied by the search engine proper, or may be applied by a ‘display engine’ once a set of documents has been retrieved by the search engine.

The searcher should be able to keep track of the documents that they have seen. If and when actual copies of the documents are available it would be useful to be able to highlight passages, and add annotations for future reference.

4.2.2.1 By zootagger

Filters can be designed to select or prioritise the contribution of individual classifiers. The selection of data sources may be an efficient way of filtering unwanted noise.

A searcher may choose to select classification that has been performed by a particular group: whether restricted to a specific geographical region, to patent examiners, classifiers from a particular organization, or perhaps just themselves.

4.2.2.2 By zootags

Selection and/or prioritisation can be performed on the basis of a near-hit in a particular dimension or dimensions.

This is discussed in more depth in relation to the search engine functionality.

4.2.2.3 By bibliographic data or document type

Selection or prioritisation on specific dates of publication, or between time periods, on authors, titles and other publication data needs to be made possible.

Similar criteria are the particular type of document, for example those that are not copyright protected, the language of the document or maybe those for which a translation is freely available.

4.2.3 Managing the credibility of information
'Four candles or fork handles?' (The Two Ronnies)

It may appear reasonable that 'obvious errors' should be correctable by Codekeepers or Sages. But then not all errors will be obvious. It may be that the original classification could be improved, or maybe further refined.

But it may be that an error is no more than a 'difference in appreciation', or it may be that an error is in fact the correct information.

If a Codekeeper or Sage believes a zootagging to be incorrect they could send a query to the zootagger in question. However this would likely represent a considerable investment.

Maybe the zootagging being accompanied by a text entry explaining why the particular zootags have been attributed would reduce such an overhead.

A pragmatic solution may be to keep the original zootags in the database, and to allow other classifiers to add both their appreciation of the original zootagging, and possible provide an alternative zootagging.

Zootaggers 'liking' a specific zootagging could be the crowd's way of lending credibility to the information. An extension of this idea is 'Collaborative Text Classification' [19] which is described as a categorization method where the content is represented by the feedback of a large number of users.

Such feedback can also be fed into a credibility rating of the zootagger.

Such 'expertise classification' systems exist across the web, for example to identify 'top reviewers' in Amazon, and are described in the literature [20].

Such credibility ratings could be factored into the search or display engines to further select and prioritise search. A searcher could select all 5 star zootaggers regardless of origin; zootagging that is perceived by the crowd as less credible could be preferred less, and ranked lower in search as a result.

It would not be efficient for a same document with alternate zootagging to be viewed multiple times during a search.

It would be possible for a search query to hit the multiple zootagging of a same document. Or it would be possible for a search query to hit only one zootagging of a same document.

One option would be to display all of the zootagging by all the zootaggers of a particular document together with the credibility ratings whenever a document is first displayed.

It may be helpful for an original zootagger to be able to select notification of the zootagging of a document they have previously zootagged. The additional zootagging information may be for aspects not identified in the original classification.

It is hoped that such a protocol will keep zootaggers on board, whilst providing a searcher with the means to take into account the credibility of information.

5. More design and functionality

5.1 Zootag evolution

Technology evolves over time, and classification schemes follow.

A root structure is rigid and does not evolve easily. Whilst it may be relatively simple to grow a new root, a reorganization of existing roots is a very complex business. How to provide for the convergence of technologies for example? What stays on each root, what goes? It is more than a simple prune-and-graft process.

A historical legacy within the IPC has resulted in a taxonomy of technology that may be different from that which an engineer or scientist might expect.

The architecture of Durham Zoo, designed around the ‘shades-of-grey’ and free from the constraints of a root structure, is easier to adapt. Editing the ZSD’s: whether by ‘dragging and dropping’ existing zootags, refining a scheme by adding new zootags, or reclassifying existing zootags to a different scheme is relatively straightforward.

Below an example of how Durham Zoo can account for a change of technology such as ‘digital convergence’.

The example relates to a Personal Digital Assistant (PDA).

A PDA of the 1980’s was a handheld device, typically with calendar, calculator, address book and note-taking functionalities. Later models included mobile phone technology.

Today a modern smartphone may have much, if not all, of the early PDA functionality. And it is likely to have music-playing and camera functionalities too. Personal music players now have Wi-Fi connections that can browse the web, and the idea of a hand-held computer without networking capability is maybe already a strange one.

This ‘digital convergence’ has blurred the boundaries between personal computing, telephone and personal music players. Such changes need to be reflected by changes in the inferences between entities on the ZSD’s.

As an example, below is a simplified ZSD of a PDA in the year 1997 and in 2020 (for more details check out the Psion PDA and Nokia 9000 Communicator).

In the year 1997, long before the smartphone revolution, the PDA was closest in functionality to a hand-held computer. There was at least one PDA incorporating mobile phone technology, but this was not typical functionality of a PDA. As far as we are aware, music playing functionality was as yet not linked with the PDA.

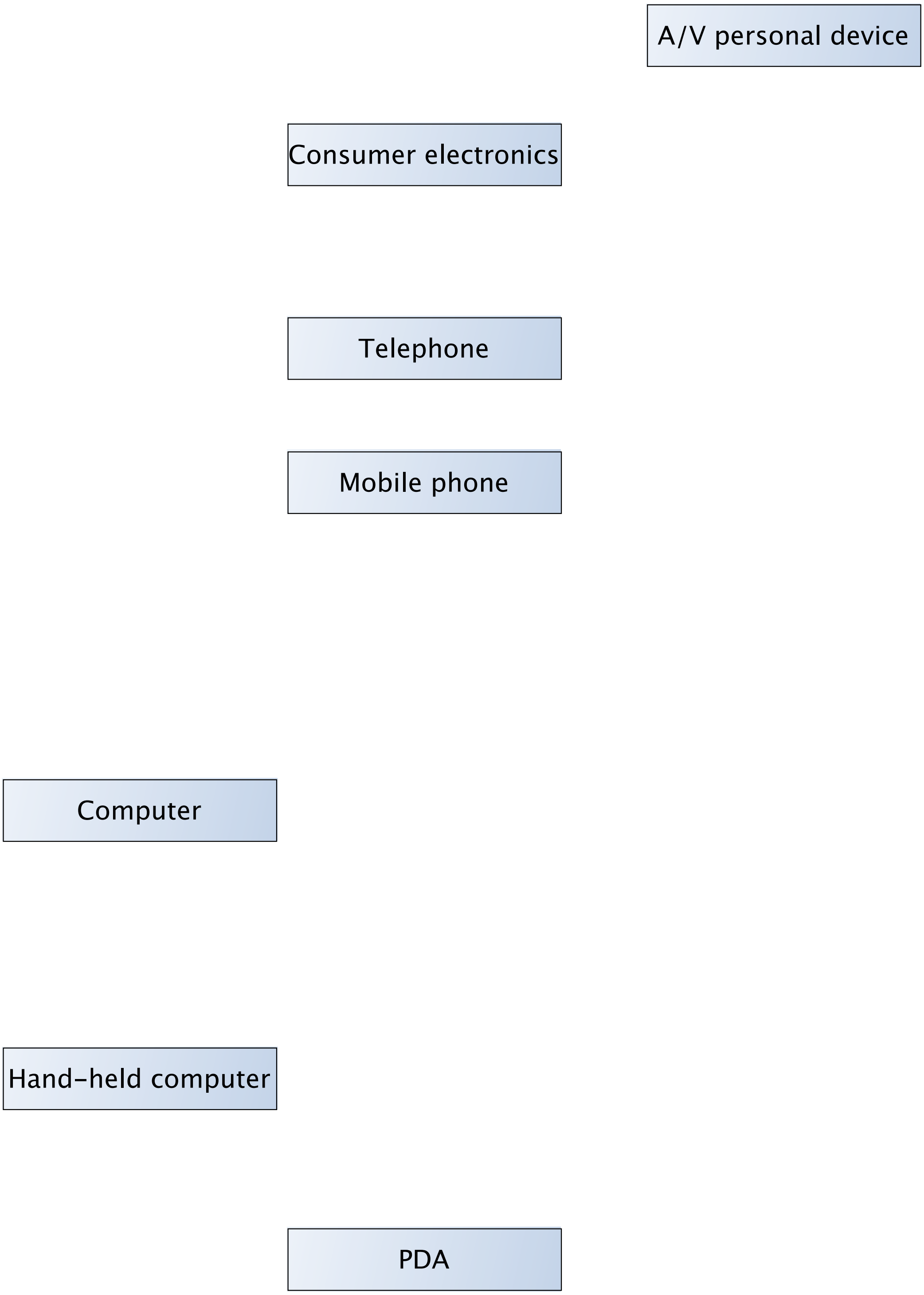

Fig. 13 PDA 1997

In 2020 the smartphones all have computing, phone and music player functionalities and so it approximates to a PDA. Similarly the personal music player has grown very PDA-like.

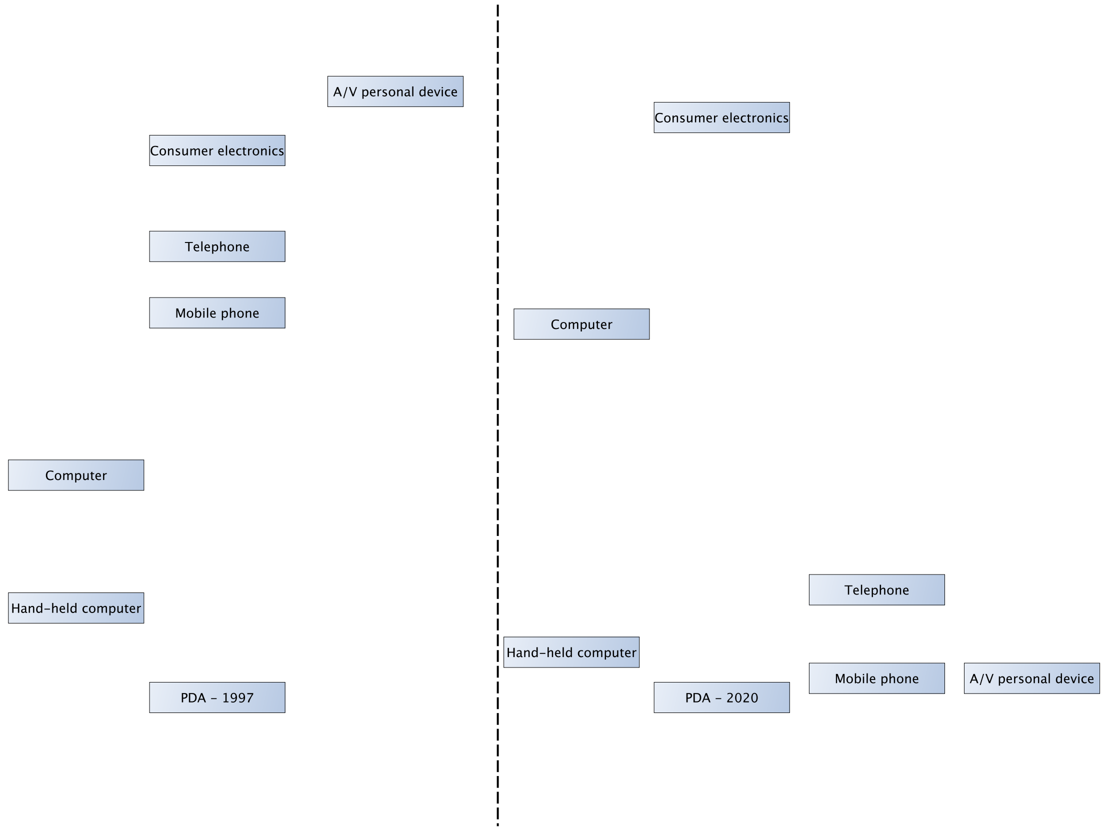

Fig. 14 PDA 1997 & 2020

To reflect the changes in technology has required no more than the shifting of the entities on the code steering diagram to alter the inference values.

In reality, by 2020, the concept of a PDA may no longer exist. Maybe a ‘smartphone’ will simply be a ‘phone’ that contains everything? Will computing power be on the cloud? Maybe the ‘phone’ will also routinely include personal health monitoring features such as blood pressure and heart rhythm? The evolution of technology needs be reflected in the ZSD's.

There may come a time when the simple shifting of entities on the ZSD's is not enough. Maybe some zootags will have been rendered obsolete. Maybe others have been rendered so similar as to be tautological.

Such a more comprehensive reorganization is achieved using 'ghost zootags' and 'zombie zootags'.

First a word about the different types of reorganization:

the reassign, e.g. A goes to B;
the split, e.g. A goes to A, A1 and A2;
the merge, e.g. A and B goes to C;
the merge and split, e.g. A and B goes to C and D.

Reorganization thus refers to changes in the zootags themselves. Reclassification refers to the updating of the zootags attributed to documents.

Although optimal in terms of the quality of the information, the reclassification of documents following a reorganisation requires time and effort. The provision of ghosts and zombies allows the reclassification effort to be completed only where necessary.

'Ghost zootags' are zootags that have been superseded, effectively 'killed off'. Ghost zootags cannot be allocated to a document, and neither do they figure in search. Ghost zootags exist uniquely to complete the information on a ZSD: providing help in understanding a present scheme or its historical perspective.

'Zombie zootags' are not living codes but neither are they completely dead. Like ghosts they cannot be allocated. However where zombies appear on a ZSD they also figure in search. Zombies can exist forever, or can be used for the transitional state between a code being a living code and a ghost.

So if a reorganisation is undertaken the 'outgoing' zootag is made a zombie. The zombie zootagged documents *do not need* to be classified to the new scheme, given that the zombie will remain on the ZSD's and continue to figure in the search process. However, where the reclassification of the zombie-zootagged documents to a newer scheme is worth the effort, this can be done. And when all the documents have been reclassified, the zombie zootag can be killed off to become a ghost.

5.2 Timewarp

Of most importance in a prior art and solution search is the here and now.

However it should be possible to store the classification scheme for each period, such that a search can be performed with the classification scheme of a particular period.

It may be difficult to travel back in time and put oneself in the mind-set of a foregone era. Having become accustomed to seeing something as standard can render an objective appreciation of the idea at the time of its inception difficult.

Once having seen the intermittent wipe functionality on a windscreen washer it is may be difficult to imagine a time when it did not exist. And it may be easy to be dismissive of the inventive nature of such an idea.

Litigation concerning a patent may benefit from an appreciation of the inference values at the time of the application. The inference values of the period concerned may better reflect the 'state of the art' and the likely assessment of the skilled person as regards 'inventive activity'.

This is the Timewarp functionality.

For this, all documents need keep their zootag history, and copies of the ZSD's maintained for each period.

Durham Zoo classification is to the most recent scheme, but with the possibility of 'morphing' back to an earlier one.

5.3 Response to new technologies

Entropy is a measure of disorder, or more precisely unpredictability.

Classifying a radically new technology may prove problematic: the classification scheme existing at the emergence of the new technology may have no provision for it.

So whilst it may be relatively simple to understand the subject matter it may be difficult to know where it goes.

Furthermore there may be few 'skilled persons' to perform the classification.

And thus the entropy in classification may be greater for an emerging technology, reducing with time as the prior art and expertise grows.
However any delay in response to a new technology will occur at exactly the time when the IP 'land grab' may be the most critical: the comparison of a patent application with the prior art forms the basis of defining the scope of conferred.

A new technology may overlap two or more existing technologies, or it may be radically new. In the cases where there is an overlap there is a danger that different disclosures relating to the new technology are classified with zootags from different fields. This is a problem, creating the incomplete and overlapping collections that can result from digital convergence.

If it does not fit well with any existing technologies it may be an 'underlap', receiving no classification in one or more dimensions.

New technologies need to be identified, zootags created and ZSD's designed as soon as is possible.

And whilst it may be that the patent offices that see the new technologies first, Durham Zoo should make every effort to react as quickly.

5.4 Reduced classification effort

The PDA ZSD demonstrates how Durham Zoo can further improve on a root-based classification scheme.

On a root-based scheme the technical knowledge and classification knowledge can be local to the individual roots. Classification may require the circulation of the same document to multiple classifiers. Phenomena such as digital convergence may require the circulation of an increased number of documents.

However in Durham Zoo digital convergence can be a force for good. In Durham Zoo the classification of a document with a single zootag effectively classifies the document with all the zootags on the ZSD's to the degree that they are applicable. The attribution of a PDA zootag, with the inference to computing, telephony and music players may preclude further circulation to the experts to the individual fields.

And whilst we wouldn't go as far as calling it 'automatic' classification', there is an element of automatic classification inherent in the scheme.

5.5 Lowering the language barrier

The user interface could be provided in any number of languages; however the zootags would have the same meaning in the different languages, this analogous with the IPC.

Zootagging could be performed in the mother tongue of the classifier. Only those documents retrieved in a search would need to be translated for further evaluation.

This has to be a better option than translating documents into a common language and searching in that one language, or even translating a search query and performing multiple language searches in parallel.

5.6 Integrating classification completed thus far

Any new scheme should, where possible, build on the massive amount of classification completed thus far, obviously subject to the agreement of the owners of the different classification.

Our initial idea was to develop correspondence tables between the zootags and other existing schemes. We foresaw one-to-one correspondence, correspondence via Boolean or fuzzy equation, we wondered to what degree artificial intelligence and machine learning could be implemented.

The combination of disparate classification schemes is discussed in patent US 2007/0136221 A1 [21].

However we preferred a human solution. And given that ‘simplicity is the ultimate sophistication’ (Leonardo da Vinci): we sought a simpler solution.

That solution is to add the existing classification information onto the different zootag diagrams. We don’t have a working system, and so for the moment we just call it our ‘grand unifying classification theory’.

As an example, using our horse ZSD, the following information could be added: an existing Japanese Patent Office classification code for a horse which matches entirely with the zootag definition, an existing American Patent Office classification code for a horse which includes the zootag definition of a pony as well as a horse, and so is weighted less than 1, and an independent British classification code for a pantomime horse which is about as light as possible.

Adding a French code for ‘Crazy Horse’ would probably take the idea too far. However, what of a rocking horse? Could not the design of a saddle for a children’s rocking horse be of interest to a saddle designer proper?

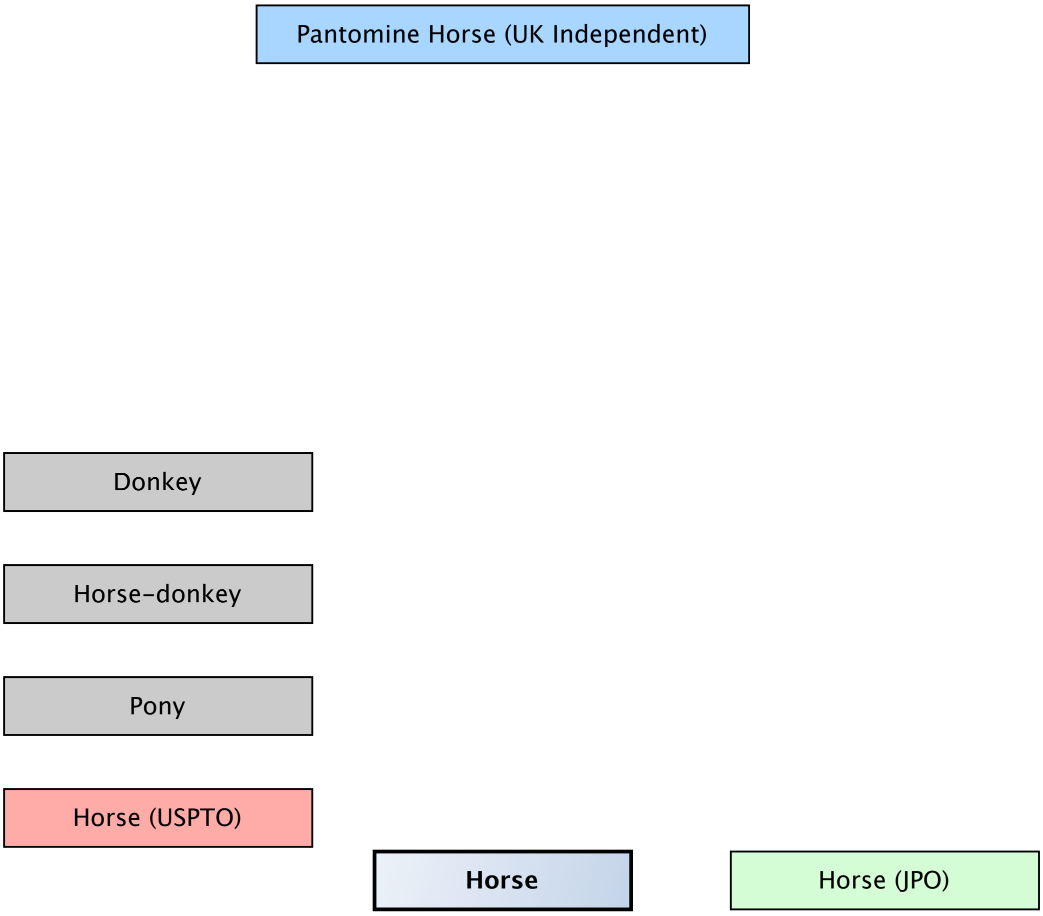

Fig. 15 Integrating classification schemes

5.7 Synergy in search: Classification And Text – the CAT Search

The zootag-associated text includes all synonyms, and a description in plain language of the zootag. As such the zootags could form the basis of a keyword search in the text of a document or set of documents.

This would also be a means for retrieving prior art documents that have not yet been zootagged.

The search in text could be performed on a set of documents retrieved using a prior zootag search, or the inverse.

Alternately, a zootag Classification And Text Search, possibly with advanced natural language processing, could produce a combined ranking of pertinent documents.

If ever the CAT search sees the light of day, we would dearly like permission to use a picture, easily found on the Internet, that features a mask-disguised cat together with two raccoons (try searching in images with 'cat and raccoons')

6.The business/busyness case

We don't believe the project should be run like a business, however we've tried to think about the project from a business-like perspective.

We came up with three questions: the could, the would and the should.

Could it work? Do we have the technology?
Would it work? We need to put information in to get information out. Could we mobilise a crowd?
Should it be tried? Is it worth the effort to try?

6.1 Could

The project uses established technology and could be made to work.

6.2 Would

'The only real wisdom is in knowing that in a collaborative environment there is most likely someone else who will know' (Socrates v 2.0).
'Better a little which is well done, than a great deal imperfectly' (Plato).

The project would only work if people were prepared to invest in it.
We believe in the power of the crowd and we believe that are enough people out there who would enjoy Durham Zoo.

There are legions of technically interested persons who dispense technical help on Internet forums for nothing more than the pleasure of exercising their expertise.

There are many that write reviews offering their expert opinion, whether on Amazon or elsewhere.

Galaxy Zoo is a great example of crowd participation for a worthwhile cause.

Publishers may be attracted to use the zootagging system as a means to better disseminate their articles via more hits and more downloads? Zootagging by an esteemed classifier may provide additional information over and above that written in a freely available abstract. This information may encourage a download of a fully copy of an article requiring payment.

Authors and inventors may be inclined to zootag their own work, again to gain greater exposure or to broaden interest in their ideas. Although not intended as a database of the documents themselves Durham Zoo could incorporate a notary for literature as well as zootags.

Likewise companies may choose to classify their company literature if they believed they could effectively publicize their products in such a manner.

Solution search may encourage botanists and zoologists to participate with nature's solutions and thus promote biomimicry.

Zootag status could even feature on LinkedIn?

6.3 Should

We believe Durham Zoo can contribute to promoting innovation, reducing duplication in research, and maybe even promoting collaboration in research.

The patent system is under pressure. The public is party to the contract that is a patent. Is it not right and proper that the public is given every opportunity to contribute in a meaningful way to the patenting process, and especially now that technology makes this a practical possibility?

We believe that zootagging could contribute to the patenting effort, raising quality and potentially reducing the numbers? The patent system is not a numbers game, but it would appear that a patent portfolio is often required to compete on equal terms. This was cannot have been the intention of the patent system [22].

But neither can it have been the intention for the holders of dubious patents to be able to hound commercially active companies with threats of litigation and the offer to buy their silence [23].

The considerations as to 'distinguishing information', and 'breadth of applicability' of Durham Zoo are pretty close to the 'inventive concept' and 'scope of protection' of a patent.

If Durham Zoo led to an improved understanding of the patent system it may improve the general public's ability to contribute to it.

A search in Durham Zoo could contribute to the process of deciding whether a patent application should be granted or not. This 'upstream-of-grant' contribution may be more efficient than trying to destroy granted patents?

However Durham Zoo could also be used for experts to provide additional prior art against granted patents, perhaps via links in the Durham Zoo database.
Durham Zoo could also provide a voice for the crowd. 'Like or dislike' style rating could be the means for the crowd to indicate whether it considers a patent trivial, or unjustifiably broad. This could eventually become a patent court of virtual instance.

Should not synergies with Wikipatents be investigated? And what of the Electronic Frontier Foundation's patent busting project and Mozilla's 'prior art initiative'? It would be good to pool resources [24] [25].

Many would argue that the concept of offering an exclusion right in return for a contribution to the art should apply generally, and not be conditional upon whether an invention relates to computer software or not.

Is it not trivial patents or erroneously granted patents, whether in software or anything else, that creates the real problems?

One way for the open-source software lobby to contribute to the prevention of such patents would be by filing defensive publications in a notary, and zootagging them.

The underlying concept of computer code may be effectively obfuscated and difficult to search. Would not the classification of software concepts by the crowd improve the quality of software patents? What do you think Richard Stallman?

We would suggest that the database be free for searchers. Advertising revenue could be generated to pay for the maintenance and development of the system. Any additional monies could fund research such as climate-change mitigation technologies or other good causes.

And whilst we believe in the not-for-profit model and originally considered a classification-for-free design, we would like to raise the possibility of a reward of some kind to classifiers, and even to the authors of literature in the database? We believe in a model that primarily rewards knowledge creation.

We would not like the information to be crawled into commercial search engines for shareholder benefit, even if it did provide an awesome user experience.

6.4 Risk analysis

'The chief danger in life is that you take too many precautions' (Alfred Adler)

Whilst we have received much positive feedback to our ideas, we have also received feedback that the system would be open to attack from all sorts of anarchistic groups, that there would be political complications, disagreements between classifiers expressing different academic viewpoints, dubious classification, wrong classification and malicious classification.

And yes, of course all of these will happen.

However that does not translate to the problems being intractable, or a reason for not giving it a go.

7. References Part 2

(Internet links valid 31.03.2012 (some need a 'return' in the browser)

1. International Patent Classification
http://www.wipo.int/ipcpub/ - refresh=page

2. Google patents
http://www.google.com/patents

3. European Patent Office Espacenet
http://www.epo.org/searching/free/espacenet.html

4. WO 2010147673 A2, Tan Weihong et al. (23-12-2010)
http://worldwide.espacenet.com/publicationDetails/biblio?DB=worldwide.espacenet.com&II=0&ND=3&adjacent=true&locale=en_EP&FT=D&date=20101223&CC=WO&NR=2010147673A2&KC=A2

5. US 2006033674 A1, J R Jr Essig (2006-02-16)
http://worldwide.espacenet.com/publicationDetails/biblio?DB=worldwide.espacenet.com&II=1&ND=3&adjacent=true&locale=en_EP&FT=D&date=20060216&CC=US&NR=2006033674A1&KC=A1

6. F-term (patent law)
http://en.wikipedia.org/wiki/F-term_(patent_law)

7. B82Y codes
http://www.wipo.int/ipcpub/ - refresh=page&notion=scheme&version=20110101&symbol=B82Y

8. Y02 codes
http://worldwide.espacenet.com/eclasrch?locale=en_EP&classification=ecla&ECLA=Y02

9. Natural language search in the IPC
http://www.wipo.int/tacsy/

10. Common Hybrid Classification
http://www.fiveipoffices.org/projects/commonhyb.html

11.The Common Hybrid Classification Project of the Five IP Offices
http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=131197

12. Cooperation Patent Classification
http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=163677

13. Combined CPC/FI Introduction into the IPC
http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=155497

14. The "Donald Duck as prior art" case, A. Engelfriet, 2006
http://www.iusmentis.com/patents/priorart/donaldduck/

15. Traditional knowledge
http://en.wikipedia.org/wiki/Traditional_knowledge

16. Controlled manipulation of nanoparticles with an atomic force microscope, T. Junno et al.,1995
http://apl.aip.org/resource/1/applab/v66/i26/p3627_s1?isAuthorized=no

17. Post-It Notes Were Invented By Accident, DAVEN, 2011
http://www.todayifoundout.com/index.php/2011/11/post-it-notes-were-invented-by-accident/

18. Dependency graph
http://en.wikipedia.org/wiki/Dependency_graph

19. Semi-Supervised Collaborative Text Classification, Rong Jin et al., 2007
http://www.cs.cmu.edu/~rahuls/pub/ecml2007-rahuls.pdf

20. Expertise Classification: Collaborative Classification vs. Automatic Extraction, T. Bogers et al., 2006
http://ilk.uvt.nl/~toine/publications/bogers.2006.sigcr2006-paper.pdf

21. Patent US 2007/0136221 A1, Sweeney et al. (14-06-2007)
http://www.google.com/patents/US20070136221.pdf

22. Google Buys Motorola: The Patent Wars Ramp Up, A Wagner, 2011
http://www.huffingtonpost.com/2011/08/15/google-motorola-mobility_n_927670.html

23. Patent troll
http://en.wikipedia.org/wiki/Patent_troll

24. PATENT BUSTING PROJECT
https://w2.eff.org/patent/

25. EFF Report: Patents and the Public Domain
http://www.securinginnovation.com/2008/06/articles/prior-art-database/eff-report-patents-and-the-public-domain/