The Ontology: Chimaera or Pegasus

Christopher Brewster, José Iria, Fabio Ciravegna and Yorick Wilks

Department of Computer Science, University of Sheffield,
211 Portabello Street, Sheffield, S1 4DP, U.K.
C.Brewster@dcs.shef.ac.uk

Abstract

In the context of the needs of the Semantic Web and Knowledge Management, we consider what the requirements are of ontologies. The ontology as an artifact of knowledge representation is in danger of becoming a Chimera. We present a series of facts concerning the foundations on which automated ontology construction must build. We discuss a number of different functions that an ontology seeks to fulfill, and also a wish list of ideal functions. Our objective is to stimulate discussion as to the real requirements of ontology engineering and take the view that only a selective and restricted set of requirements will enable the beast to fly.

1. Introduction

It has been widely accepted that one of the major challenges for the Semantic Web is to efficiently and effectively construct ontologies (McGuiness, 2003). How this will be done is a function of many different factors. Just as every bridge built by a civil engineer is different depending on the width of the river, the ground on each side, the size of the bridge and a myriad of other factors, so in constructing ontologies there are many ways to set about it each with their own consequences.

Artificial Intelligence has a long tradition of narrowing down problems to such an extent that while on the one hand they become possible to implement, the solutions are often of value only in a limited sphere. This is true whether we look at microworlds or expert systems. This is exactly one of the challenges for which ontologies are supposed to be a partial solution. Ontologies are the core representational medium for knowledge in the Semantic Web. However, ontologies have been sold to the research community as a panacea, a remedy for all things for all men. For example, we find the word “ontology” being used to describe taxonomies such as those of Yahoo, lexical databases such as WordNet, and logically coherent constructs over which reasoning systems can operate (like FaCT, Racer, etc. Horrocks 1998, Haarslev and Möller 2001)

Rather than going over the somewhat philosophical issues as to what an ontology *is* or rather what it *ought* to be, in this paper we will enumerate a number of potential requirements and consider their ramifications. We will argue that not all of them can be satisfied at the same time, because in effect we are dealing with a beast that, like the mythological Chimera of the Ancient World, is expected to be a multiplicity of animals (cf Figure 1). Such a beast is unreal.

To continue our metaphorical excursion, we can conceive of the task of building ontologies from the perspective of civil engineering. We can consider what our foundations are and what materials we have to build with. We can consider the intended construction and its contradictory requirements. I will leave it as an exercise for the reader to develop the necessary tools to take us from foundations to delivery.

Figure 1: The Chimera of Arezzo

2. The Foundations

The geology of knowledge is in flux An ontology is a model, a representation of knowledge. However, that knowledge which we try to model is in a continuous state of change and thus in many cases (depending on the scope of the ontology) the model is out of date from the moment of completion. This ever more true the larger the size of an ontology, and consequently we see large hand built ontologies continuously being revised (e.g. the Gene ontology Gene Ontology Consortium 2005), often with quite radical changes. If our purpose were merely to capture an image, or record the state of knowledge at a given time, it would be a less complex task. But what we want is to represent knowledge so as to do things with that representation so there needs to be a level of analytical detail much like an outstanding architectural model of a planned building.

When one says knowledge is in flux this means a number of things. At a simple level new concepts or new ideas are added to the body of knowledge. At a more complex level, new ways of analysing existing concepts or items in the world arise. In the 20th century four models of knowledge were articulated which have implications on how we view ontologies. First, there is the traditional positivist view of knowledge, where each discovery is added to the overall construct. In this view of knowledge, an ontology would be monolithic and change would only occur at the edges as new items are added. There would be no need for internal change. Secondly, in the Popperian view of knowledge (Popper, 1959), an ontology would reflect a theory about the world which would in effect be tested continuously so that whenever it was proved wrong a new theory (in our case a new ontology) would need to be constructed. A third view was that of Thomas Kuhn, who believed that theories of knowledge fit into 'paradigm' or 'disciplinary matrix' reflecting the common consensus at a given time (Kuhn, 1962). As new discoveries are made so there is a fraying at the edges until eventually a revolution is triggered an an entirely new theory has to be provided. Models of knowledge from different paradigms are 'incomesurate' which means there is no way to compare them or evaluate one with respect to another because the meaning of basic terms or concepts is affected by the theories they are embedded in. Finally there is the view of Quine who believed that knowledge is like a 'force field' and as things change at the edges so this has a knock on effect throughout the system sometimes triggering minor sometimes major changes in the internal structure (Quine, 1951). In such a case an ontology would need to be designed to be continuously restructured in its internal form depending on the changes along the periphery.

Clearly our traditional conception of ontologies appears to fit the positivist monolithic approach. Certainly such is the case of ontologies like Cyc (Lenat et al., 1994). The problem for ontology construction is that one of the other theories is certainly closer to the real progression of human knowledge, although it is doubtful whether these philosophical perspectives actually capture the real complexity.

Human labour is expensive. Excavating knowledge by hand from the substrate of human artifacts (whether a collection of minds or texts) is extremely costly. It is costly in time, it is costly in financial terms and it is very error prone. It has consequently been a foundational tenet of ontology building that we must automate the task ideally in its entirety. The reality has been that the extent of automation has been proportionate to the triviality of the knowledge acquired by software systems. Thus for example a system like *Knowitall* can identify that a *biologist* is a subclass of *scientist*[REF]. So can a human being, faster, more easily and cheaply. What *Knowitall* can claim is to finding many instances of *cities* which are not present in a standard gazetteer. The issue is to identify where human labour is necessary and where it is most effective.

Knowledge is abstract, text is concrete. We need to remember that we cannot read human minds. We cannot directly draw on the accumulated knowledge of either an individual or a community. An ontology as a form of knowledge representation is an abstract model of what we individually or collectively believe is true about the world and the actors in it. Our only concrete source of information in this regard is text in that we can analyse texts on a computer, manipulate them and extract certain types of information from them.

Our dependence on text has its advantages and disadvantages. On the positive side, we have now over a generation of research in text analysis and computational linguistic methods to draw on and exploit. There are large quantities of text whether in specific corpora or on the Internet. The existence of the Internet, however, provides both a challenge and an opportunity in that while it is an almost infinite resource, it is very repetitive, unreliable and ambiguous. When what is needed is an ontology of a given domain, what is *not* needed are texts using the same terms in an entirely different domain. There is also the problem that in effect the visible web is only the tip of the iceberg and we would like ways of accessing the knowledge available in the Deep Web. Surely a great deal of relevant knowledge is stored in its deeper manifestations. This is all the more relevant when we consider the next foundational issue.

The implicit nature of text. The greatest problem with extracting knowledge from text is that a large proportion of what we need to identify is not in fact explicitly stated in the text. As we have argued elsewhere (Brewster et al., 2003), a text is an act of knowledge maintenance. In writing a text, each and every author assumes a shared set of concepts, ideas and terms of reference. This is inevitable otherwise communication could not arise. A given text interacts in specific ways with this assumed ontology. First, it re-enforces the assumptions of that background knowledge by telling the reader which ontology to use to process the text. Secondly, the text alters the links, associations and instantiations of concepts already present in the ontology. Thus a primary purpose of a text at some level is to change the relationship between existing concepts, or change the instantiations of those concepts. Finally, a text may affect a domain ontology by adding new concepts to the existing domain ontology.

In each case, however, a text takes the background knowledge, in effect the background ontology for granted. Texts rarely make explicit statements concerning the ontology. In Quinean terms, texts alter our knowledge of the periphery; they do not make explicit statements about the core of knowledge. This may seem strange but it is inevitable given the nature of communication. The engineering consequences are substantial however:

1. It means we cannot expect from a given domain specific corpus of texts to derive the underlying ontology because at nearly every step of the way that ontology is assumed.

2. It means we need to find techniques round the tacit manner in which knowledge is expressed.

3. We need to establish coherent methods for going outside the corpus to identify the missing knowledge.

4. The search for explicit knowledge must be focussed on that which is actually needed, i.e. no need to relearn basic ontologies.

3. The Construct

We can identify certain specific roles that this strange modern beast, an ontology, is required to play. In enumerating these, no attempt is made to claim that they are compatible with each other and we believe in fact that this is a significant issue as to which requirements should take priority in any given construction effort. We begin first with a brief wish list[1] and then move on to certain functions of ontologies.

[1] *Pace* Donald Rumsfeld http://en.wikipedia.org/wiki/Rumsfeld

3.1. The Wish List

Ontologies should represent things we know we know. Most ontologies of various shapes and sizes which have been created up to now try to represent the knowledge that we all think we know we know. One aspect of this is that for any given community an ontology is supposed to represent its "share conceptualisation" (Gruber, 1993) so inevitably an ontology reflects shared agreement, what Kuhn would have called the 'paradigm'. Thus we find that *Cyc* has attempted to capture commonsense knowledge i.e. what we all agree about (Lenat et al., 1994). Similarly the editors of the Gene Ontology wish to create "consistent descriptions" for their field. This is only possible if there is wide agreement on what is known and what it is called. These efforts are eeriely reminiscent of the (unsuccessful) seekers for a universal language in the late medieval and renaissance time (Rossi, 2000).

There has been an unfortunate tendency in ontology engineering to develop quite complex software systems which produce relatively trivial output (cf. Navigli and Velardi 2004, Cimiano et al. 2004, Etzioni et al. 2004). Complex algorithms are written, resources are trawled, and many texts analysed to determine that, for example, a **ferry service** *isa* **boat service**. The question arises whether this is the foundation of something much larger or whether these systems will never be able to break out of limited domains. If the latter, where does the bottleneck lie?

Another issue is the extent to which ontologies should represent only that which is thoroughly understood. Given the logicist tradition in AI, and its effect through Description Logics on ontology engineering, there is a danger that our representation of the world will only reflect that which is understood sufficiently from a logical perspective (cf.below).

Ontologies should make accessible things we know we don't know. Some of the major intended tasks for ontologies both in the arena of the Semantic Web and Knowledge Management is to act as vehicle for knowledge retrieval (Schraefel et al., 2004). This is an explicit aim in the Gene Ontology and clearly this is the most obvious function of ontologies/taxonomies such as Yahoo or the Open Directory (`www.dmoz.org`). Possibly the most effective demonstration of the potential of using a form of ontology is to be found in the work of Hearst on the Flamenco project where great emphasis has been placed on allowing the metadata to facilitate exploratory search (Hearst et al., 2002; Yee et al., 2003).

Another manner in which ontologies allow us, or should allow us to access things we know we don't know is in ontology population. This is in effect a form of Information Extraction (Ciravegna and Wilks, 2003) and Etzioni et al. (2004) give a typical successful example in finding a multitude os cities absent from a standard gazetteer. However, there is another sense that an ontology must make knowledge accessible which is to allow the identification of relations between objects which we did not know existed. Let us imagine we know about subject x and subject y but so far we have been unaware of any impact, relationship of effect of one on the other. An ontology should provide a means of representing or inferring this type of knowledge. The real issue is where this sort of knowledge comes from, and given that everything is potentially related to everything else, how this knowledge can be evaluated as of significance.

Ontologies should make accessible things we don't know we don't know. The greatest challenge lies in identifying knowledge which we do not know we do not know. This is a major practical problem in many cases where there is an open ended continuous input of textual data. For example, there are people employed by the Environment Agency (UK) whose job it is to identify new threats to the environment. In effect, they do not know what they do not know, because threats to the environment can arise from a multitude of origins which are unpredictable. Similar situations arise in economic analysis or security intelligence.

It is an open question whether we can construct systems which will allow the spotting of unknown unknowns. We can conceptualise this as shown in Figure 2. There is an existing ontology or body of knowledge (the known known). There is an unknown ontology (it may be unknown for example because until now it has been considered irrelevant) - this is the unknown unknown. There is a connection, an impact, an effect of items in the unknown ontology on the known.

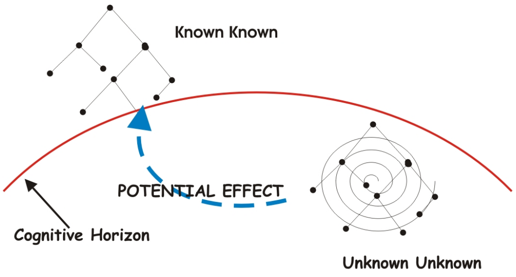

Figure 2: Conceptualising the unknown unknown

Knowledge is normally acquired by following a thread, one thing leads to another. The fundamental question to ask here is whether this is the only way to discover knowledge (in our practical context) or whether we can envisage other more creative and dynamic means.

3.2. Functions of an Ontology

A simplistic view would say that an ontology is a model of the world which can be used to reason about it. However, there is a whole range of functions, assumptions and aspirations encoded in a given type or instance of an ontology. Following Davis et al. (1993), we consider five functions, or rather let us say five animals which an ontology tries to embody.

An ontology is a surrogate. Intelligent entities go through processes of reasoning about the world, often in order to plan actions. The reasoning involves internal representations but the objects reasoned about are usually external, outside in the world. Consequently the ontology or knowledge representation is surrogate standing for the objects outside in the world. The correspondence between the surrogate or knowledge representation is the semantics. The 'fidelity' of the correspondence depends on what the knowledge representation captures from the real thing and what it omits. Perfect fidelity is impossible. Two important consequences are that every ontology must unavoidably lie even if it is merely by omission, and that all forms of reasoning about the world will reach false conclusions at some stage. The soundness of the reasoning process cannot avoid this because the representation is in some way incorrect. It is essentially a practical engineering decision to find ways minimising the the errors given the specific task at hand.

There are further consequences. Firstly one of the major claims for ontologies is that they will facilitate the interchange of knowledge between (for example) agents, or the reuse in different systems. However, if each agent or system has an imperfect model of its universe, knowledge interchange or sharing may increase or compound errors which were not visible in the initial use of an ontology. Second, and closely related, ontologies of the same domain will inevitably model different aspects of the external world depending on the focus and assumptions of the ontology's authors.

A Ontology is a set of Ontological Commitments. This may appear tautological but the choice of ontology is also a "decision about how and what to see in the world" (Davis et al., 1993, :19). This is both unavoidable because representations are imperfect and useful because it allows the representation to focus on that which the representation's author considers relevant or interesting. They see these choices as allowing us to cope with the overwhelming complexity and detail of the world. Consequently it is the content of the representation i.e. the set of concepts chosen and their inter-relation which provides a particular perspective on the world. The choice of notation (logic, LISP, or OWL) is unimportant.

It is interesting that, with respect to ontologies, an immense amount of effort has been expended in developing and defining ontology representation languages, and in contrast relatively little effort has been made to analyse what ontological commitments particular ontologies make. The only exception to this has been Guarino's critique of structures such as WordNet for not conforming to a logician's world view in terms of consistency and logical rigour (Guarino 1998; Gangemi et al. 2001 but cf. Wilks 2002).

An inherent assumption of the all authors in this field is that 'concepts' are the key building blocks, and we manipulate concepts with words. All ontologies I have encountered use words to represent the concepts and to mediate or provide a correspondence with the external world. Consequently a large range of items in the world or experiences which do not lend themselves readily to verbal expression cannot be modelled. We could describe this as the 'Ontological Whorf-Sapir Hypothesis' i.e. that that which cannot be captured by words cannot be represented in an ontology.

An Ontology or Knowledge Representation is a Fragmentary Theory of Intelligent Reasoning. The way a knowledge representation is conceived reflects a particular insight or understanding of how people reason[2]. The selecting any of the currently available representation technologies (such as logic, frames, knowledge bases, or connectionism) commits one to fundamental views on the nature of intelligent reasoning and consequently very different goals and definitions of success.

The OWL language which has been developed by the W3C consortium as a standard language for describing ontologies for the Semantic Web comes in three flavours (`http://www.w3.org/2004/OWL/`). The existence of these three flavours reflects the differing traditions that have merged in the current effort to construct a standard. This does not diminish the point that by choosing a specific form of knowledge representation, there is a commitment to specific views about the nature of intelligence. After all Minsky himself in his original paper on frames noted that his work was a "partial theory of thinking" (Minsky, 1975) and so equally there are limits and presumptions in a standard like OWL which have not been fully spelled out but of which the 'flavours' are an indication. Minsky himself has taken a very humble view, stating in an interview, "I want AI researchers to appreciate that there is no one 'best' way to represent knowledge. Each kind of problem requires appropriate types of thinking and reasoning – and appropriate kinds of representations" (Minsky and Laske, 1992). This comment should raise many questions as to the longer term viability of a standardisation approach like that of the Semantic Web community with OWL and the successful use of a limited range of knowledge representations.

An Ontology is a Medium for Efficient Computation In the final analysis an ontology must allow for computational processing, and consequently issues of computational efficiency will inevitably arise. For example, using taxonomic hierarchies both "suggests taxonomic reasoning and facilitates its execution" (Davis et al., 1993, :27).

Clearly the development of the different flavours of OWL are a recognition of this fact, but in seeking sufficient speed severe restrictions on the reasoning capacity of the representation have had to be made. More generally, it could be noted that since all ontologies depend on a propositional view of knowledge in order to begin to be computationally tractable, already a very restricted view of what it is possible to represent has arisen. The fact that *OWL Full* is not guaranteed to be 'decidable' unfortunately does not guarantee it to be sufficiently powerful to represent the whole gamut of what we can consider to be knowledge.

An Ontology is a Medium of Human Expression All forms of knowledge representation including ontologies are both mediums of expression for human being *and* ways for us to communicate with machines in order to tell them about the world. That knowledge representation is a form of human expression is something frequently forgotten in the field. In Nirenburg and Wilks (2001), for example, Nirenburg's insistence on the possibility of precise unambiguous meaning in a 'representational language' ignores this fact. Wilks' response that the symbols in a representation are fundamentally language-like essentially reflects that a representation language *is* a means of human communication with all its dynamism, ambiguity and extensibility. This fact is frequently forgotten in the ontology building

[2] The very use of a *representation* is a commitment to symbolic AI or what Haugeland call "Good Old Fashioned AI" (Haugeland, 1985, :112ff.).

communities (such as the Semantic Web) who vainly believe that their ontologies will achieve ‘precision’ and ‘exactness’ in the meaning of the terms (classes, concepts, etc.) in their ontologies.

Figure 3: Pegasus on an Attic vase

4. Conclusion

In this paper, we have laid out a series of conditions or state of affairs concerning the construction of ontologies. We have argued that the foundations and resources available for their construction are highly problematic. We believe that the construct to be achieved has many contradictory requirements both in what it should do and roles it plays. All these cannot be achieved simultaneously, and there has to be a degree of compromise in the design and execution of the construction task. We need to seek to construct a Pegasus and not a Chimera, otherwise it will not fly (cf. Figure 3).

It is inappropriate to pretend that an ontology can be all things to all people and this has been the underlying rhetoric in recent years even if not explicitly stated. In conclusion, we hope that these points will stimulate discussion and perhaps more importantly guide the kind of tools and methodologies to be developed in the future.

5. Acknowledgements

This work is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and The Open University. It is also supported by the EPSRC funded ABRAXAS project (http://nlp.shef.ac.uk/abraxas/) on “Automating Ontology Learning for the Semantic Web”.

References

Brewster, Christopher, Fabio Ciravegna, and Yorick Wilks, 2003. Background and foreground knowledge in dynamic ontology construction. In *Proceedings of the Semantic Web Workshop, Toronto, August 2003*. SIGIR.

Cimiano, Philipp, Andreas Hotho, and Steffen Staab, 2004. Comparing conceptual, divise and agglomerative clustering for learning taxonomies from text. In R. López de Mántaras and L. Saitta (eds.), *ECAI 2004 Proceedings of the 16th European Conference on Artificial Intelligence, 22 - 27 August, Valencia, Spain*. IOS Press.

Ciravegna, Fabio and Yorick Wilks, 2003. Designing adaptive information extraction for the semantic web in amilcare. In Siegfried Handschuh and Steffen Staab (eds.), *Annotation for the Semantic Web*, Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam.

Davis, Randall, Howard Shrobe, and Peter Szolovits, 1993. What is a knowledge representation. *AI Magazine*, 14(1):17–33.

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, , and Alexander Yates, 2004. Methods for domain-independent information extraction from the web: An experimental comparison. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI 04)*. AAAI Press.

Gangemi, Aldo, Nicola Guarino, and Alessandro Oltramari, 2001. Conceptual analysis of lexical taxonomies: the case of wordnet top-level. In *Proceedings of the international conference on Formal Ontology in Information Systems*. ACM Press.

Gene Ontology Consortium, 2005. The gene ontology. http://www.geneontology.org.

Gruber, T. R., 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 6(2):199–221.

Guarino, Nicola, 1998. Some ontological principles for designing upper level lexical resources. In *Proceedings of the First International Conference on Language Resources and Evaluation, LREC 98*.

Haarslev, Volker and Ralf Möller, 2001. Description of the racer system and its applications. In Carole A. Goble, Deborah L. McGuinness, Ralf Möller, and Peter F. Patel-Schneider (eds.), *Working Notes of the 2001 International Description Logics Workshop (DL-2001), Stanford, CA, USA, August 1-3, 2001*, volume 49 of *CEUR Workshop Proceedings*.

Haugeland, John, 1985. *Artificial Intelligence: The Very Idea, 1985*. MIT Press.

Hearst, Marti, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee, 2002. Finding the flow in web site search. *Commun. ACM*, 45(9):42–49.

Horrocks, I., 1998. Using an expressive description logic: Fact or fiction? In *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR’98), Trento, Italy, June 2-5, 1998*. Morgan Kaufmann.

Kuhn, Thomas, 1962. *The Structure of Scientific Revolutions*. Chicago, Illinois: University of Chicago Press.

Lenat, Douglas B., Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd, 1994. Cyc: Toward programs with common sense. Technical report, MCC and Stanford.

McGuiness, Deborah L., 2003. Ontologies come of age. In Dieter Fensel, James Hendler, Henry Lieberman, and Wolfgang Wahlster (eds.), *Spinning the Semantic Web*, chapter 6. Cambridge, MA: MIT Press, pages 171–196.

Minsky, Marvin, 1975. A framework for representing knowledge. In P. Winston (ed.), *The Psychology of Computer Vision*. McGraw-Hill, New York, pages 211–277.

Minsky, Marvin and Otto Laske, 1992. A conversation with marvin minsky. *AI Mag.*, 13(3):31–45.

Navigli, Roberto and Paula Velardi, 2004. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, 30(2).

Nirenburg, Sergei and Yorick Wilks, 2001. What's in a symbol: Ontology, representation, and language. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(1):9–23.

Popper, K. R., 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Quine, Willard V., 1951. Two dogmas of empiricism. *The Philosophical Review*, 60:20–43. Reprinted in W.V.O. Quine, From a Logical Point of View (Harvard University Press, 1953; second, revised, edition 1961.

Rossi, Paolo, 2000. *Logic and the Art of Memory: The Quest for a Universal Language*. London: Athlone Press. Tr. from Italian by Stephen Clucas. Original edition "Clavis Universalis: Arti della Memoria e Logica combinatori da Lullo a Leibniz", Mulino, Bologna, 1983.

Schraefel, M. C., Nigel R. Shadbolt, Nicholas Gibbins, Stephen Harris, and Hugh Glaser, 2004. Cs aktive space: representing computer science in the semantic web. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*. ACM Press.

Wilks, Yorick, 2002. Ontotherapy: or how to stop worrying about what there is. Invited presentation, Ontolex 2002, Workshop on Ontologies and Lexical Knowledge Bases, 27th May. Held in conjunction with the Third International Conference on Language Resources and Evaluation - LREC02, 29-31 May, Las Palmas, Canary Islands.

Yee, Ka-Ping, Kirsten Swearingen, Kevin Li, and Marti Hearst, 2003. Faceted metadata for image search and browsing. In *Proceedings of the conference on Human factors in computing systems*. ACM Press.