Dynamic Reconciliation of Semantics among Numerous Local Ontologies: A Preliminary Report

Michael N. Huhns and Larry M. Stephens
Center for Information Technology
University of South Carolina
Columbia, SC 29208
{huhns,stephens}@sc.edu

May 8, 2001

Abstract

This report describes a methodology by which information from large numbers of independently developed sources can be associated, organized, and merged. The central hypothesis is that a multiplicity of ontology fragments, representing the semantics of the independent sources, can be related to each other automatically *without* the use of a global ontology. That is, any pair of ontologies can be related indirectly through a *semantic bridge* consisting of many other previously unrelated ontologies, even when there is no way to determine a direct relationship between them. The relationships among the ontology fragments indicate the relationships among the sources, enabling the source information to be categorized and organized. A preliminary evaluation of the methodology has been conducted by relating 53 small, independently developed ontologies for a single domain. A nice feature of the methodology is that common parts of the ontologies reinforce each other, while unique parts are de-emphasized. The result is a *consensus* ontology.

1. Introduction

The research reported herein targets the following basic problem: a Web search will typically yield a large number of pointers to Web sites—some of which are relevant and some of which are irrelevant; the sites might be ranked, but they are otherwise unorganized, and there are too many for a user to investigate manually. The problem is familiar and many solutions have been proposed. The solutions range from requiring the user to be more precise in specifying search criteria, to constructing more intelligent search engines, to requiring Web sites to be more precise in describing their contents. A common theme for all of the approaches is the use of ontologies for describing both requirements and sources. Unfortunately, ontologies are not a panacea unless everyone adheres to the same one, which does not yet exist in a comprehensive enough form (in spite of attempts, such as the Cyc Project, to create one). Moreover, even if one did exist, it is unlikely that it would be adhered to, considering the dynamic and eclectic nature of the Web.

To overcome this, there are three possible approaches by which information from large numbers of independently developed sources can be associated, organized, and merged semantically: (1) each Web site will use the same terminology with agreed-upon semantics (a method considered improbable), (2) each Web site will use its own terminology, but provide translations to a global ontology (a method considered difficult, and thus unlikely), and (3) the methodology described herein based on small, local ontologies. Our methodology relies on Web sources that have been annotated with ontological information [Pierre 2000], which is consistent with several visions for the semantic Web [Berners-Lee 1999; Heflin and Hendler 2000; Berners-Lee, Hendler, and Lassila 2001]. The sources and ontologies must be for similar domains—else there would be no interesting relationships among them—but they will

undoubtedly have dissimilar formulations and terminology because they will have been developed independently.

Our central hypothesis is that a multiplicity of ontology fragments, representing the semantics of the independent sources, can be related to each other automatically *without* the use of a global ontology. That is, any pair of ontologies can be related indirectly through a *semantic bridge* consisting of many other previously unrelated ontologies, even when there is no way to determine a direct relationship between them. Rather than scale being a problem, additional ontologies can make it easier—or even possible—to relate two ontologies. The resultant merged ontologies provide a semantic characterization of the set of sources and their domain, and gives the apparent effect of a large single ontology serving as a global hub for interactions. The methodology produces a means for agents and other information system components to interoperate.

We are evaluating the methodology by applying it to a large number of independently constructed ontologies in a restricted domain. The success or difficulty encountered in this effort will constitute useful scientific knowledge of benefit to others working in ontology development and heterogeneous information integration.

2. Semantic Reconciliation of Separately Developed Ontologies

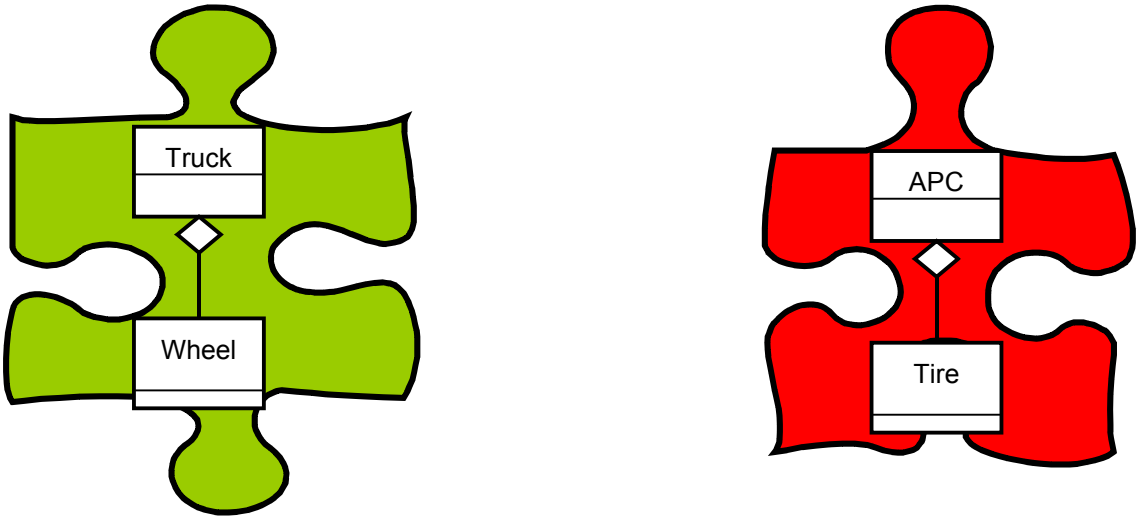
In agent-assisted information retrieval, a user will describe a need to his agent, which will translate the description into a set of requests, using terms from the user's local ontology. The agent will contact on-line brokers and request their help in locating sources that can satisfy the requests. The agents must reconcile their semantics in order to communicate about the request. This will be seemingly impossible if their ontologies share no concepts. However, if their ontologies share concepts with a third ontology, then the third ontology might provide a "semantic bridge" to relate all three. Note that the agents do not have to relate their entire ontologies, only the portions needed to respond to the request.

The difficulty in establishing a bridge will depend on the semantic distance between the concepts, and on the number of ontologies that comprise the bridge. The methodology we are investigating is appropriate when there are large numbers of small ontologies—the situation we expect to occur in large and complex information environments. Our metaphor is that a small ontology is like a piece of a jigsaw puzzle, as depicted in Figure 1. It is difficult to relate two random pieces of a jigsaw puzzle until they are constrained by other puzzle pieces. We expect the same to be true for ontologies.

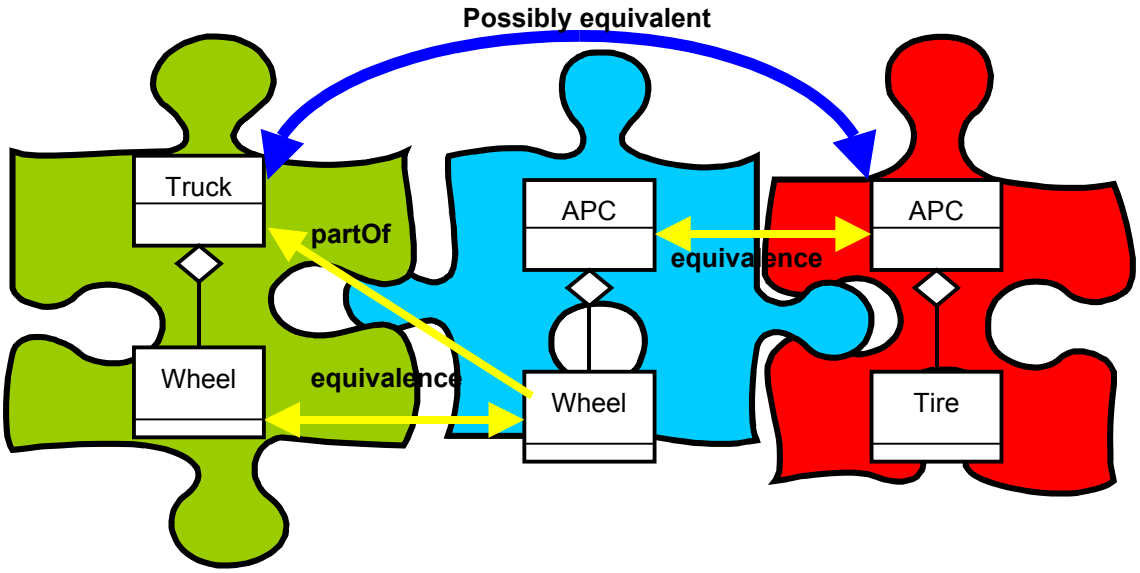
Analysis

Two concepts can have the following seven mutually exclusive relationships between them: *subclass*, *superclass*, *equivalence*, *partOf*, *hasPart, sibling*, or *other*. If a request contains three concepts, for example, and the request must be related to an ontology containing 10 concepts, then there are 7 x 3 x 10 = 210 possible relationships among them. Only 30 of the 210 will be correct, because each of the three concepts in the request will have exactly one relationship with each of the 10 concepts in the source's ontology. The correct ones will be determined automatically by applying constraints among the concepts within an ontology, and constraints that arise from discovered constraints among multiple ontologies. Once the correct relationships have been determined, the major ones of interest are *equivalence* and *sibling* or, where those do not exist, the most specific *superclass* or most specific *partOf*.

Consider the example in Figure 1. The ontology fragment on the left would be represented as *partOf(Wheel, Truck),* while the one on the right would be represented as *partOf(Tire, APC)*. There are no obvious equivalences between these two fragments. The concept *Truck* in the first ontology could be related to *APC* in the second by *equivalence, partOf, hasPart, subclass, superclass,* or *other*. There is no way to decide which is correct. Now consider the addition of the middle ontology fragment *partOf(Wheel, APC)*. With this added information, there is evidence that we could link as *equivalent* the concepts *Truck* and *APC*, and *Wheel* and *Tire*.



(a) Two ontology fragments with no obvious relationships between them



(b) The introduction of a third ontology reveals equivalences between components of the original two ontology fragments

Figure 1. Ontologies can be made to relate to each other like pieces of a jigsaw puzzle

This example exploits the existence of the relation *partOf*, which is common to all three ontologies. Other domain-independent relations, such as *subclassOf*, *instanceOf*, and *subrelationOf*[1], will be necessary for the reconciliation process. Moreover, the following properties of relations are needed for relating occurrences of the relations to each other [Stephens 1991]: reflexivity, symmetry, asymmetry, transitivity, irreflexivity, and antisymmetry. Domain concepts and relations can be related to each other by converse/inverse, composition, (exhaustive) partition, part-whole (with 6 subtypes), and There must be

[1] Examples of subrelations are (1) *on* is a subrelation of *above* in spatial relations, (2) *daughterOf* is a subrelation of *childOf* in familial relations, and (3) *cityLocation* is a subrelation of *countryLocation* in geographic relations.

some minimum set of these fundamental relations that are understood and used by all local ontologies and information system components.

In attempting to relate two ontologies, a system might be unable to find correspondences between concepts because there might not be enough constraints and similarity among their terms. However, trying to find correspondences with other ontologies might yield enough constraints to relate the original two ontologies. As more ontologies are related, there will be more constraints among the terms of any pair of ontologies. In this way, the presence of many small ontologies becomes an advantage. It is also a disadvantage in that some of the constraints might be in conflict. We will make use of the preponderance of evidence to resolve these statistically.

3. Experimental Methodology

We asked a group of 53 graduate students in computer science, who were novices in constructing ontologies, to each construct a small ontology for the Humans/People/Persons domain. The ontologies were written in DAML and were required to contain at least 8 classes with at least 4 levels of subclasses; a sample ontology is shown in Figure 2.

Using a string-matching algorithm and other heuristics (see the Appendix for precise characterizations of the heuristics), we constructed a single merged ontology from the 53 component ontologies. The component ontologies described 864 classes. After applying our algorithm, the merged ontology contained 281 classes in a single graph with a root node of the DAML #Thing. It related all of the concepts from these ontologies, leaving no orphan concepts, i.e., there was some relationship (path) between any pair of the 281 concepts.

Next, we constructed a *consensus ontology* by counting the number of times classes and subclass links appeared in the component ontologies when we performed the merging operation. For example, the class Person, and all similar classes whose name matched using our string-matching algorithm, appeared 14 times. The subclass link from Mammals (and its matches) to Humans (and its matches) appeared 9 times. We termed these values the "reinforcement" of a concept.

Redundant subclass links were removed and the corresponding transitive closure links were reinforced. That is, if C has subclass A with reinforcement 2, C has subclass B reinforced *m* times, and B has subclass A reinforced *n* times, then the link from C directly to A was removed and the remaining link reinforcements were increased by 2.

We then removed from the merged ontology any classes or links that were not reinforced by appearing multiple times. The result, shown in Figure 4, represents an implicit consensus among the ontology writers about what concepts should appear in the Humans/People/Persons domain and how they should be related.

Finally, we applied an *equivalence heuristic* for collapsing classes that have common reinforced superclasses and subclasses. For example, Figure 4 contains both Humans and Person. The equivalence heuristic found that all reinforced subclasses of Person are also reinforced subclasses of Humans, and all reinforced superclasses of Person are also reinforced superclasses of Humans. It thus deems that Humans and Person are the same concept. This heuristic is similar to an inexact graph matching technique such as [Manocha et al., 2001]. The collapsed consensus ontology, now containing 36 classes related by 62 subclass links, is shown in Figure 5.

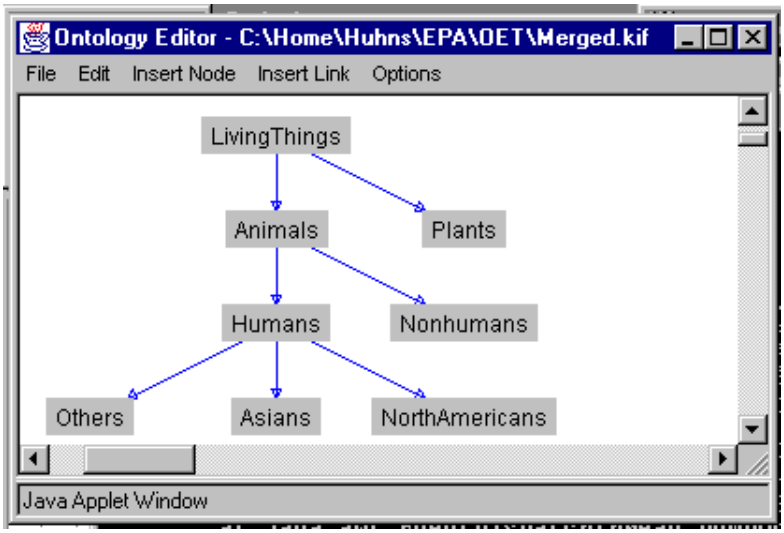


Figure 2. A typical small ontology used to characterize a Web site about people (all links denote subclasses)

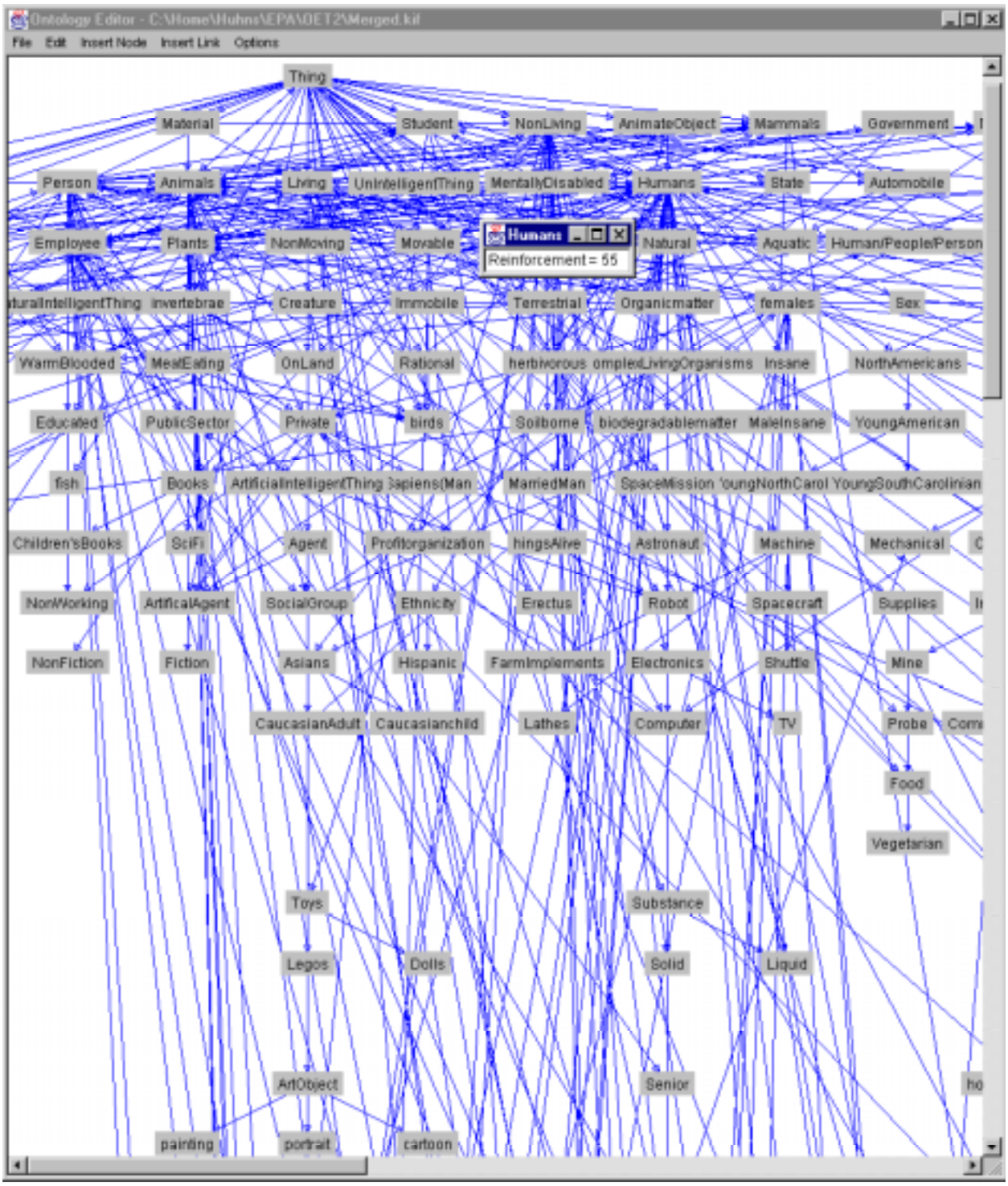


Figure 3. A portion of the ontology formed by merging 53 independently constructed ontologies for the domain Humans/People/Persons. The entire ontology has 281 concepts related by 554 subclass links

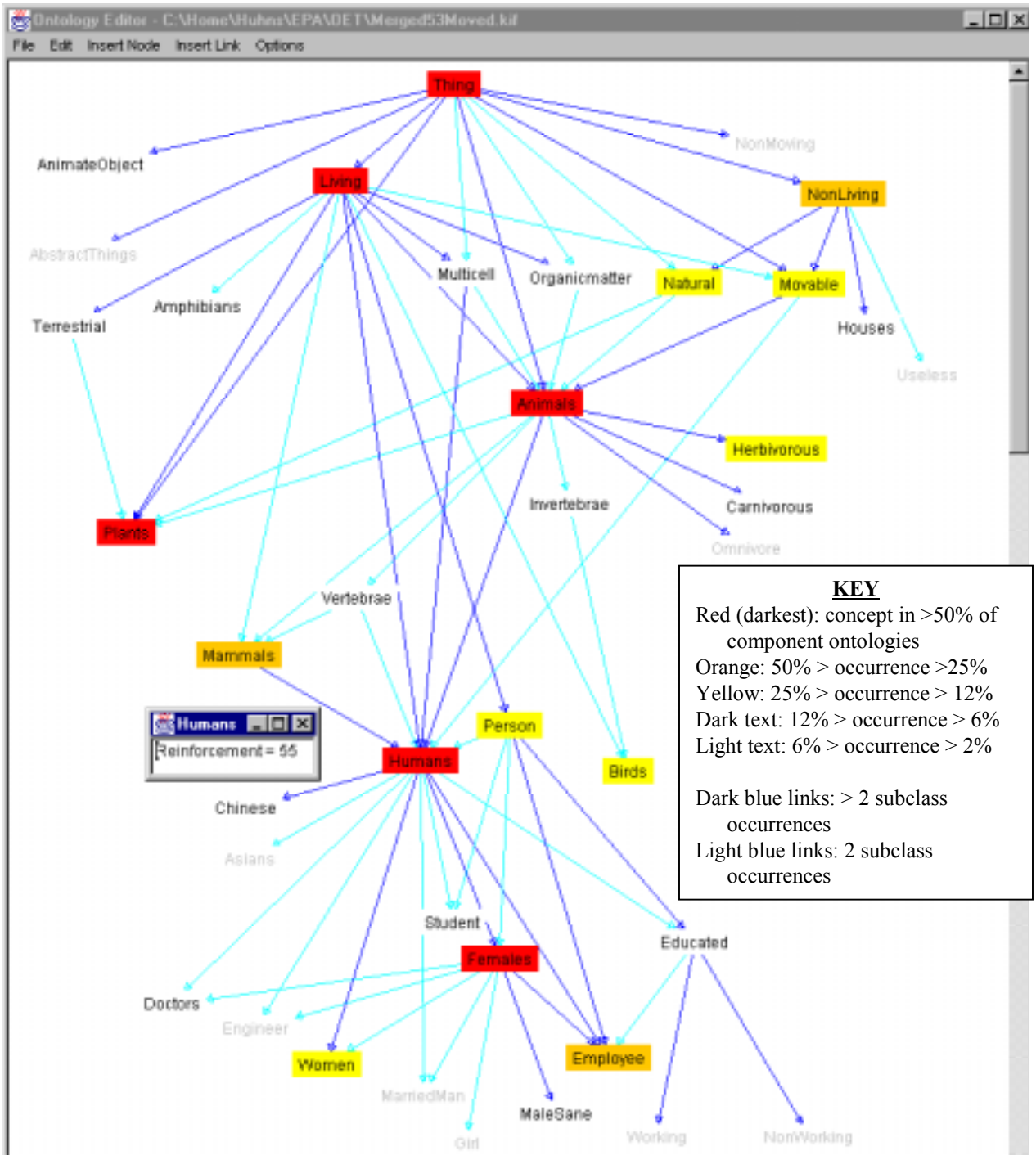


Figure 4. The consensus ontology formed from a merger of 53 ontologies independently constructed for the domain of Humans/People/Persons. There are 38 concepts with 71 subclass links that appear in more than one of the 53 original ontologies

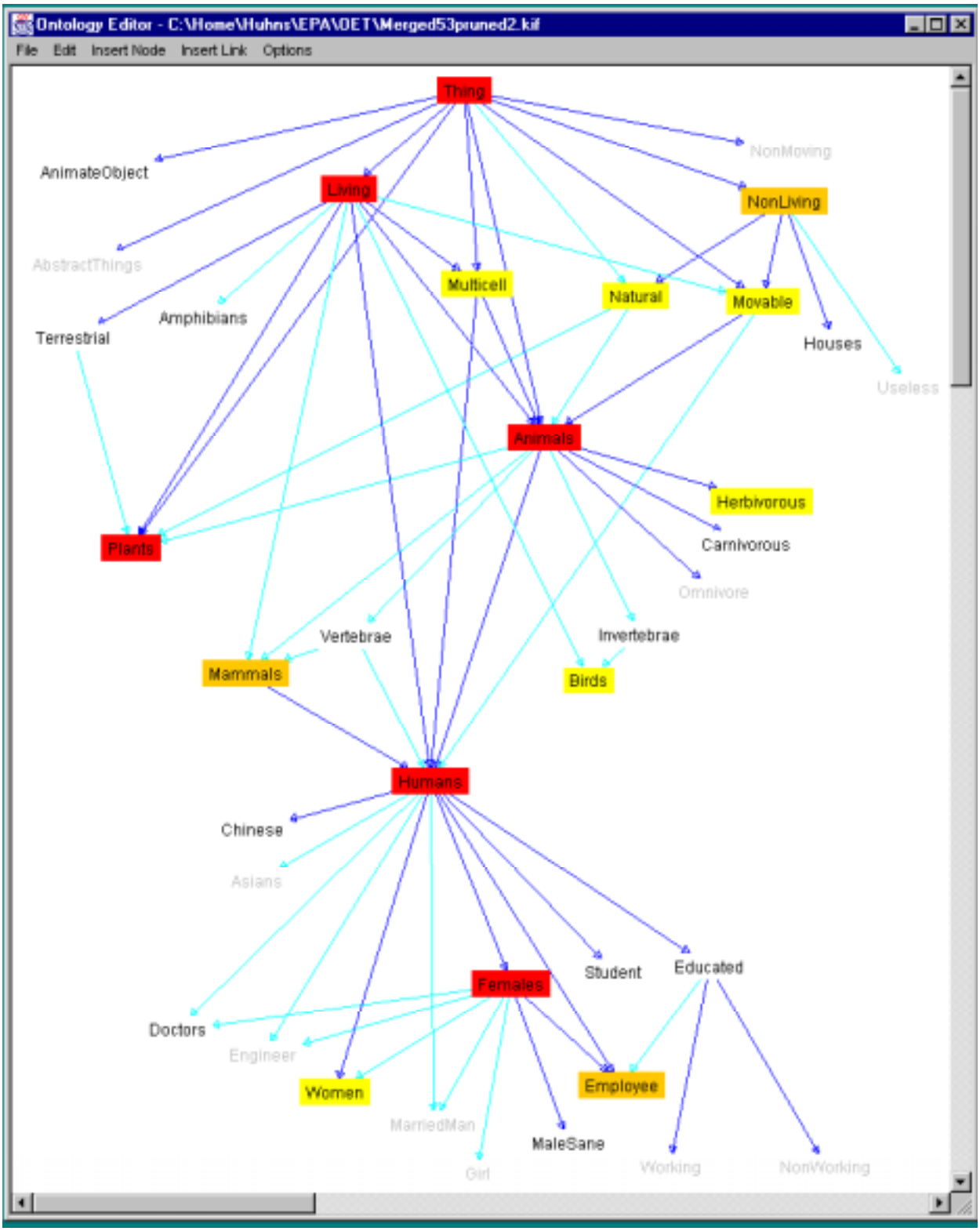


Figure 5. The final consensus ontology formed by merging concepts with common subclasses and superclasses. The resultant ontology contains 36 concepts related by 62 subclass links.

4. Discussion of Results

A consensus ontology is perhaps the most useful for information retrieval by humans, because it represents the way most people view the world and its information. For example, if most people wrongly believe that crocodiles are a kind of mammal, then most people would find it easier to locate information about crocodiles if it were located in a mammals grouping, rather than where it factually belonged.

Although promising, our experiments and analysis so far are very preliminary. We used the following simplifications:

- Our string-matching algorithm did not use morphological analysis to separate the root word from its prefixes and suffixes, and did not identify negated concepts, such as "uneducated" versus "educated."
- We did not use synonym information, such as is readily available from WordNet, and so did not for example merge "meat eating" and "carnivorous."
- We did not make use of any properties of the classes, as is done in terminological subsumption.
- We used only subclass-superclass information, and have not yet made use of other important relationships, notably *part-of*.

Our future research will systematically address each of these limitations and will use the more sophisticated merging algorithm found in the Appendix.

5. Semantic Distance

The information retrieval measures of precision and recall are based on some degree of match between a request and a response. The length of a semantic bridge between two concepts can provide an alternative measure of conceptual distance and an improved notion for relevance of information. Previous measures relied on the number of properties shared by two concepts within the same ontology, or the number of links separating two concepts within the same ontology. These measures not only require a common ontology, but also do not take into account the density or paucity of information about a concept. Our suggested measure does not require a common ontology and is sensitive to the amount of information available in the domain.

6. Ongoing Research

Our hypothesis, that a multiplicity of ontology fragments can be related to each other automatically without the use of a global ontology, appears promising, but our investigation of it is just beginning. We are proceeding according to the following plan:

- Utilize our graduate students to produce domain-specific ontologies that contain additional attributes and relationships, such as part-whole.
- Improve the algorithm for relating ontologies described in section 2. The algorithm will be based on methods for partial and inexact matching, and will make extensive use of common ontological primitives, such as *subclass* and *partOf*. The algorithm will take as input ontology fragments and produce mappings among the concepts represented in the fragments. The algorithm will control the computational complexity of its ontology-relating operation by making use of constraints among known ontological primitives.
- Evaluate the computational complexity of the algorithm both theoretically and experimentally.
- Develop metrics for successful relations among ontologies, based on the number of concepts correctly related, as well as the number incorrectly matched. The *quality* of a match will be based on semantic distance, as measured by the number of intervening semantic bridges.

The results of our effort will be (1) software components for semantic reconciliation, and (2) a scientific understanding of automated semantic reconciliation among disparate information sources.

7. Conclusion

Imagine that in response to a request for information about a particular topic, a user receives pointers to more than 1000 documents, which might or might not be relevant. The technology developed by our research would yield an organization of the received information, with the semantics of each document reconciled. Our goal is to enable users to retrieve dynamically generated information that is tailored to their individual needs and preferences.

Our premise is that it is easier for individuals or small groups to develop their own ontologies, whether or not a global one is available, and that these can be automatically and *ex post facto* related. We are working to determine the efficacy of local annotation for Web sources, as well as the ability to perform reconciliation that is qualified by measures of semantic distance. The success of this research will enable software agents to resolve the semantic misconceptions that inhibit successful interoperation with other agents and that limit the effectiveness of searching distributed information sources.

8. Bibliography

[Ambite et al.] Jose Luis Ambite, Yigal Arens, Eduard Hovy, A. Philpot, L. Gravano, V. Hatzivassiloglou, and J. Klavens, "Simplifying Data Access: The Energy Data Collection Project," *IEEE Computer*, February 2001, pp. 47-54.

[Berners-Lee et al. 2001] Tim Berners-Lee, James Hendler, and Ora Lassila, "The Semantic Web," *Scientific American*, May 2001.

[Chaffin and Herrmann 1988] R. Chaffin and D. Herrmann, "The nature of semantic relations: a comparison of two approaches," in *Relational Models of the Lexicon: Representing knowledge in semantic networks,* M. W. Evens, ed., Cambridge University Press, Cambridge, England, 1988, pp. 289-334.

[Delugach 1993] Harry S. Delugach, "An Exploration Into Semantic Distance," *Lecture Notes in Artificial Intelligence*, No.754, pp. 119-124, Springer-Verlag, Berlin, 1993.

[Foo et al. 1992] Norman Foo, Brian J. Garner, Anand Rao, and Eric Tsui, "Semantic Distance in Conceptual Graphs," in *Current Directions in Conceptual Structure Research,* ed. L. Gerhotz, pp. 149-54, Ellis Horwood, 1992.

[Heflin and Hendler 2000] Jeff Heflin and James Hendler, "Dynamic Ontologies on the Web," *Proc. 17th National Conference on Artificial Intelligence (AAAI-2000)*, AAAI Press, 2000.

[Heflin et al. 1998] Jeff Heflin, James Hendler, and Sean Luke, "Reading Between the Lines: Using SHOE to Discover Implicit Knowledge from the Web," in *AAAI-98 Workshop on AI and Information Integration*, 1998.

[Huhns and Stephens 1989] M. N. Huhns and L. M. Stephens, "Plausible Inferencing Using Extended Composition," *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, IJCAI-89,* Detroit, MI, August 1989, pp. 1420-1425.

[Lassila and Swick 1999] Ora Lassila and Ralph R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," W3C Recommendation, 1999. http://www.w3.org/TR/REC-rdf-syntax/

[Luke et al. 1997] Sean Luke, Lee Spector, David Rager, and James Hendler, "Ontology-based Web Agents," in *Proceedings of First International Conference on Autonomous Agents,* 1997.

[Mahalingam and Huhns 1997] K. Mahalingam and M. N. Huhns, "An Ontology Tool for Distributed Information Environments," *IEEE Computer,* vol. 30, no. 6, pp.80-83, June 1997.

[Manocha et al., 2001] Nitish Manocha, Diane J. Cook, and Lawrence B. Holder, "Structural Web Search Using a Graph-Based Discovery System," *ACM Intelligence*, vol. 12, no. 1, Spring 2001, pp. 20-29.

[Pierre 2000] John M. Pierre, "Practical Issues for Automated Categorization of Web Sites," *Electronic Proc. ECDL 2000 Workshop on the Semantic Web*, Lisbon, Portugal, September 2000. http://www.ics.forth.gr/proj/isst/SemWeb/program.html

[Stephens and Chen 1996] Larry M. Stephens and Yufeng F. Chen. “Principles for Organizing Semantic Relations in Large Knowledge Bases,” *IEEE Transactions on Knowledge and Data Engineering*, 8(3), June 1996, pp. 492-496.

[Wiederhold 1994] Gio Wiederhold, “An Algebra for Ontology Composition,” *Proc. 1994 Monterey Workshop on Formal Methods*, U.S. Naval Postgraduate School, 1994, pp. 56-62.

9. Appendix: Heuristics for Merging Component Ontologies

Using the relations of Section 2.1, our methodology is embodied in the following algorithm (similar to one we developed for plausible inferencing among Cyc relationships [Huhns and Stephens 1989]):

Given RDF ontologies *A* and *B* (both based on the RDF Schema specification) having nodes $n_A(i)$, $i=1,2,\ldots,N$ and $n_B(k)$, $k=1,2,\ldots,M$ and relationship arcs $r_A(i1,i2)$ and $r_B(k1,k2)$,

- Perform string matching among $n_A(i)$ and $n_B(k)$, $\forall$ i,k, to determine candidate matches
- Perform synonym matching among $n_A(i)$ and $n_B(k)$, $\forall$ i,k, to determine additional candidate matches
- Discard matches where $n_A(i1)$ matches $n_B(k1)$ and $n_A(i2)$ matches $n_B(k2)$ but where $r_A(i1,i2)$ is inconsistent with $r_B(k1,k2)$; matches that remain are presumed to represent the relation *equivalence*[2]
- Add additional relations:
  - **If** $n_A(i) \equiv n_B(k)$
    $\wedge$ $n_A(j)$ is a *subclass*[3] *(superclass/hasPart/partOf)* of $n_A(i)$
    then $n_A(j)$ is a *subclass (superclass/hasPart/partOf)* of $n_B(k)$
  - **If** $n_A(i1) \subseteq n_A(i2) \subseteq n_A(i3)$ $\wedge$ $n_B(k1) \subseteq n_B(k2) \subseteq n_B(k3)$
    $\wedge$ $n_A(i1) \equiv n_B(k1)$ $\wedge$ $n_A(i3) \equiv n_B(k3)$
    then the relation between $n_A(i2)$ and $n_B(k2)$ is either *sibling*, *subclass*, *superclass*, or *equivalence*
  - **If** $n_A(i1)$ *partOf* $n_A(i2)$ *partOf* $n_A(i3)$
    $\wedge$ $n_B(k1)$ *partOf* $n_B(k2)$ *partOf* $n_B(k3)$
    $\wedge$ $n_A(i1) \equiv n_B(k1)$
    $\wedge$ $n_A(i3) \equiv n_B(k3)$
    then the relation between $n_A(i2)$ and $n_B(k2)$ is either *sibling*, *partOf*, *hasPart*, or *equivalence*[4]

Considering an additional ontology *C* introduces constraints that enable relations to be added as follows:

- **If** $n_A(i1) \equiv n_C(j1)$
  $\wedge$ $n_C(j2) \equiv n_B(k2)$
  $\wedge$ $n_A(i1) \subseteq n_A(i2)$
  $\wedge$ $n_C(j1) \subseteq n_C(j2)$
  $\wedge$ there are no known relationships between $n_A(i1)$ and $n_B(k1)$
  $\wedge$ there are no known relationships between $n_A(i2)$ and $n_B(k2)$
  then the relationship between $n_A(i1)$ and $n_B(k1)$ and $n_A(i2)$ and $n_B(k2)$ is either *sibling*, *subclass*, *superclass*, or *equivalence*
- **If** $n_A(i1) \equiv n_C(j1)$
  $\wedge$ $n_C(j2) \equiv n_B(k2)$
  $\wedge$ $n_A(i1)$ *partOf* $n_A(i2)$
  $\wedge$ $n_C(j1)$ *partOf* $n_C(j2)$
  $\wedge$ there are no known relationships between $n_A(i1)$ and $n_B(k1)$
  $\wedge$ there are no known relationships between $n_A(i2)$ and $n_B(k2)$
  then the relationship between $n_A(i1)$ and $n_B(k1)$ and $n_A(i2)$ and $n_B(k2)$ cannot be *other*

---

[2] The relation *equivalence* is denoted by $\equiv$

[3] The relation *subclass* is denoted by $\subseteq$

[4] The relation cannot be subclass (superclass) because if $n_A(i2) \subseteq n_B(k2)$, then at least one of the equivalence relations $n_A(i1) \equiv n_B(k1)$ or $n_A(i3) \equiv n_B(k3)$ must instead be subclass (superclass).