

Title:

MachineProse: An Ontological Framework for Scientific Assertions

Authors:

Deendayal Dinakarbandian, MD, PhD, Yugyung Lee, PhD, Kartik Vishwanath, Rohini
Lingambhotla, MS
Department of Computer Science and Informatics, School of Computing & Engineering,
University of Missouri-Kansas City, Kansas City, Missouri, MO 64110, USA

To whom correspondence should be addressed:

Deendayal Dinakarbandian
Department of Computer Science and Informatics
School of Computing & Engineering
University of Missouri-Kansas City, Kansas City, Missouri, MO 64110, USA
Email address: dinakard@umkc.edu
Phone: (816) 235-5942
Fax: (816) 235-5159

ABSTRACT

Objective: The idea of testing a hypothesis is central to the practice of biomedical research. However, the results of testing a hypothesis are published mainly in the form of prose articles. Encoding the results as scientific assertions that are both human and machine-readable would greatly enhance the synergistic growth and dissemination of knowledge.

Design: We have developed *MachineProse* (MP), an ontological framework for the concise specification of scientific assertions. MP is based on the idea of an assertion constituting a fundamental unit of knowledge. This is in contrast to current approaches that use discrete concept terms from domain Ontologies for annotation and assertions are only inferred heuristically.

Measurements: We use illustrative examples to highlight the advantages of MachineProse over the use of Medical Subject Headings (MeSH) and keywords in indexing scientific articles.

Results: We show how MP makes it possible to carry out semantic annotation of publications that is machine-readable and allows for precise search capabilities. In addition, when used stand-alone, MP serves as a knowledge repository for emerging discoveries. A prototype for proof of concept has been developed that demonstrates the feasibility and novel benefits of MP. As part of the MP framework, we have created an Ontology of relationship types with about 100 terms optimized for the representation of scientific assertions.

Conclusion: MachineProse is a novel semantic framework that we believe may be used to summarize research findings, annotate biomedical publications and support sophisticated searches.

Keywords: Medical Subject Headings; Causality; Artificial Intelligence;

I. INTRODUCTION

This paper presents MachineProse, an ontological framework for encoding knowledge that has the potential to radically transform the way in which we report, search for and analyze the results accruing from biomedical research. We first discuss the motivation behind our work, and describe the current paradigm and its limitations for exchanging scientific findings. We then introduce the significance of a scientific assertion as a fundamental knowledge construct for semantic indexing. This is followed by the main section on MachineProse, an ontological framework based on scientific assertions, where we illustrate its benefits both in encoding and searching for knowledge. Finally, we present a case study that includes a preliminary prototype, and discuss the ramifications of this approach.

II. BACKGROUND

A. Motivation

Biological and medical research has become increasingly specialized with a growing number of subfields. Nevertheless, a large proportion of biological and medical research continues to be based on the formulation and testing of hypotheses. Following this, the transmission of ideas from a research group to the world typically consists of i) Publication of the results and ii) Access to, and assimilation of, the reported information by the community. This process is currently aided by indexed databases based on controlled vocabularies and information technology. Despite the impressive advances in computer hardware and internet technology made in the recent past, the quantity of publications and the complexity of the findings make it challenging to keep up to date with the latest discoveries, even within a narrow field of research.

Hence, there is a need to find more effective ways to convey and correlate emerging discoveries in machine readable form.

B. Current Paradigm

Publication of results from scientific research. The bulk of scientific information is published mainly in the form of research articles in prose. In addition to publication in journal articles, some data is also deposited into structured databases. For example, DNA or protein sequences are submitted to databases like Genbank (Benson et al. 2004) or SwissProt (Boeckmann et al. 2003), 3D structures to the PDB (Berman et al. 2000) and protein interaction data to BIND (Bader et al. 2003). Embedded in these submissions are nuggets of new information that accrued as a result of the research. The additional information in a paper usually serves to validate the conclusions of the research and set them in context. Most journals provide free access to their abstracts. In addition, organizations like People's Library of Science (PLOS) (Bernstein et al. 2003) and Biomed Central provide unrestricted access to all articles published by them. An increasing number of journals are joining PubMed Central (Eisen et al. 2002) to allow free access to articles published in them – immediately or with a short time lag.

Current aids to access and assimilation of published research. The current paradigm has the following salient features: i) Research findings are mostly published in the form of articles in prose, ii) Only a small percentage of the data is published at the outset into databases, iii) Manual as well as automated analysis of published literature is used to summarize findings reported in the form of prose or submitted to databases, iv) Indexing of either Text words or Ontology-based terms is used to retrieve relevant articles or entries in a database.

The Medical Subject Headings (MeSH) system (Nelson et al. 2004) for annotation of articles is widely used for the annotation of publications, particularly for MEDLINE. However, it has the

following disadvantages. The annotation of articles is not done at the source, i.e., by the respective authors who know best about what is being reported in the paper. Frequently, at the time of acceptance of an article, all that is done by way of annotation is the addition of a few keywords – drawn from a journal-specific list of terms and not necessarily the MeSH standard. The National Library of Medicine (NLM) Indexing Initiative (Aronson et al. 2000) aims to automate the process of MeSH indexing in trying to reduce the exclusive reliance on humans (Bachrach et al. 1978). However, the Medical Text Indexer (MTI) (Aronson et al. 2004) used to suggest suitable terms for indexing offers limited recall and precision (Aronson et al. 2004; A MEDLINE Indexing Experiment using Terms suggested by MTI 2002). As a result, the MTI is consulted less than 20% of the time, with less than 10% of its suggestions being accepted (Aronson et al. 2004). Therefore, the task of tagging articles/abstracts with MeSH terms falls to a team of catalogers at the National Library of Medicine (NLM), the quality of whose work represents their best effort but is suboptimal to annotation at the source. A second problem is that the MeSH terminology is fairly coarse grained, even though as many as 22,568 descriptors are specified, with more being added with each release. For example, the current version of MeSH lists only “Angioneurotic Edema” under “Urticaria,” and accepted terms in clinical practice like “Chronic Urticaria” and “Cholinergic Urticaria” are absent. A third problem, and perhaps the most important, is that Boolean combinations of search terms are inadequate to describe relationships. For example, a pair of concepts, “Immunosuppressant” & “Asthma,” could occur in a document by sheer coincidence. Firstly, the two terms might be used in the abstract or article in independent contexts. Indexes based on current Ontologies or controlled vocabularies alone cannot filter out such coincidental occurrence. Secondly, even if the article does deal with the

use of a specific immunosuppressant in asthma, annotation with MeSH does not address the nature of the relationship, e.g., was it effective or not?

Several text-mining approaches have been used to extract knowledge from published literature. (Cohen et al. 2005; Rebholz-Schuhmann et al. 2005). These range from the recognition of terms to the inference of relationships (or facts), including a few that tackle the problem of suggesting causality. The simplest approach to the recognition of concepts in literature is to look for matches between words occurring in text and entries in controlled vocabularies or Ontologies. This has the following limitations. Firstly, false negatives are likely when the language in the text is not represented in the reference vocabulary. This can be exacerbated when new terms appear as a result of recent research that are not yet part of the standard terminology. Recently, support vector machines (Mika et al. 2004) have been used for the automated recognition of protein names in abstracts that depend on the inclusion of contextual information in the input to the algorithm. Hidden Markov Models have also been successfully applied to the general problem of recognizing biomedical names, even without reference to dictionaries (Zhou et al. 2004). Secondly, when comparing against multiple vocabularies, e.g., the Unified Medical Language System (UMLS) Metathesaurus (Browne et al. 2003), word ambiguity is a problem (Aronson et al. 1997). Finally, recognition of terms that refer to predefined concepts is but the first step in extracting knowledge from text. The next level of analysis, namely extracting facts or relations from text has been the subject of considerable research (Fiszman et al. 2003; Krauthammer et al. 2002; Libbus et al. 2002; Rzhetsky et al. 2004). A simple approach would be to map phrases to the UMLS (Rindflesch et al. 1993). Other forms of natural language processing have been used to extract facts (Sneiderman et al. 1996). These rely on using a combination of syntactic and semantic information to make inferences (Libbus & Rindflesch

2002; Rindflesch et al. 2000). A layered approach combining the initial identification of semantic propositions (Srinivasan et al. 2002) with subsequent evaluation for causal relations implicating genes in disease has also been attempted (Rindflesch et al. 2003).

Approaches other than natural language processing have also been used. A high-frequency of co-occurrence of a pair of terms is indicative of a relationship between two entities. This has been exploited by using association rule mining to find putative relations that meet a minimum of support and confidence which are subsequently evaluated for matches against a set of semantic templates (Zhu et al. 2003). Another idea is that of using semantic templates to recognize relations in free text (Divoli et al. 2005). Textpresso (Muller et al. 2004) is an example of a large scale document retrieval system that is based on indexing a large corpus of publications on *C. elegans* based on an Ontology. Common to all the above approaches is the re-extraction of knowledge from biomedical publications. Though some of these approaches report a high degree of precision, most of them do so in a limited context and frequently do not evaluate recall since the number of false negatives is unknown.

Attempts have been made to formalize the representation of concepts and data in the published literature. Telemakus (Revere et al. 2004) uses terms from the UMLS MetaThesaurus to synthesize research reports that include methods and findings based on a schema. Karp et al. have developed integrated representations of metabolic pathways across several species (Krieger et al. 2004) and a knowledgebase dedicated to *E. coli* (Keseler et al. 2005). Sim et al. have developed Sysbank (Carini et al. 2003) and Trial Banks (Sim et al. 2004), a formalized representation of randomized clinical trials to aid evidence-based medicine. The Reactome (Joshi-Tope et al. 2005) is a collaborative effort to develop a curated knowledgebase of biological processes.

The next section introduces an approach that complements text-mining to alleviate the problem of limited precision and recall. Like the approaches discussed in the previous paragraph, it aims to create a knowledge-base that parallels findings published in the form of prose, but it differs in its wider applicability and has several unique advantages which are discussed below.

C. Significance of Scientific Assertions to the Biomedical Research Process

Is there scope for improving on the current paradigm described above? We argue there is because of a large amount of redundant effort in current practice. This is exemplified by the labor-intensive role of human beings in finding relevant papers in the first place. Further, once found, it takes a fair amount of effort and time to extract the required information from a paper. Ideally, if one agreed upon a formal model of representing information, machines (computer programs) could aid in the process of keeping scientists and professionals up to date.

Given a biomedical paper, let us focus on the question: “How has our knowledge of the world changed after publication of this article?” The answer to the question may be distilled into the scientific assertion(s) that the paper makes. There is usually a plethora of information in a paper but most of this serves merely to justify the assertions, and set them in context. These details are generally of peripheral importance. Similarly, at the receiving end, the assertions are what the reader registers and carries away, even after reading the full paper. At the moment, abstracts serve the role of summarizing papers. Some journals require abstracts to be organized into sections, but this is still not machine-readable as unrestricted prose is used.

On the other hand, if the conclusions of a paper were summarized in a machine-readable formal structure of assertions that is not inordinately onerous, this would greatly aid both the submission and dissemination of cutting-edge scientific information. This would amount to a semantic shortcut between expression and comprehension of scientific findings with minimal loss of

information. Such a scheme would have several exciting repercussions. Firstly, the assertions could be used to index, and therefore query, scientific publications with an unprecedented degree of precision and recall. Secondly, the database of assertions would mirror the MEDLINE database in being a “bullet-item” summary of recent research results and could *per se* be the subject of interesting analyses and data mining. For example, one could study trends in research or find efficient ways to generate guidelines for evidence-based medicine. The originality of, or support for a finding could be quickly ascertained. One could even set up polling software agents to report the moment the precise answer to a highly specific query is reported for a paper, e.g., the gene locus for a disease, or the cure for a disease that one is interested in.

Is such a scheme feasible? We posit it is and such a scheme is the subject of this paper.

MachineProse is an Ontological framework for scientific assertions that conceptualizes the domain of scientific assertions and offers numerous distinct advantages over current approaches. For proof of concept, we have chosen to focus on the area of Allergy & Immunology, but the approach is applicable to any scientific domain and has the potential to enhance the interdisciplinary exchange of scientific discoveries.

III. THE *MACHINEPROSE* MODEL

A. Scientific Assertion as a Fundamental Unit of Knowledge

We are used to thinking of science in term of hypotheses. On the flip side, we may view the result of testing a hypothesis as an assertion. A scientific assertion may be represented in its simplest form as a factual relationship between a pair of entities. For example, “Chronic autoimmune urticaria *is Caused by* anti-thyroid antibody” is an example of an assertion (Fig. 1) where the underlined terms are entities and the word in italics represents the relationship. A given paper may affirm, negate or be inconclusive about the assertion. For example, the negative

form of this assertion would be “Chronic auto-immune urticaria *is NOT Caused by* anti-thyroid antibody.” An example of an inconclusive paper with respect to an assertion is the Cochrane review by Dean *et al* (Dean *et al.* 2004). It is inconclusive with regard to the assertion “Azathioprine *is effective in the treatment of* asthma,” citing the need for further studies. Thus, any given assertion may be separated into two aspects (Table 1), the root assertion *per se*, and whether the reported research affirmed it, negated it or could not reach a conclusion either way. Frequently, a term in an assertion may consist of not a discrete drug or disease but either a combination regimen of drugs or a constellation of multiple conditions. This is modeled by use of the logical ‘AND’ condition in representing the assertion. In contrast, assertions having complex entities combined by logical ‘OR’ conditions can be represented essentially as a set of independent assertions. For example, “Aminophylline OR Theophylline *cause* bronchodilation” may be modeled as two separate assertions with each of the two drugs.

B. Advantages of MachineProse Assertions Compared to Domain Ontologies

There are several critical differences between an approach based on assertions and those based on terms from biomedical ontologies or controlled terminologies. These are described below. *Assertions confer richer semantics than term-based annotation.* In MachineProse, the “scientific assertion” used to annotate a document or data source is an entire triplet, i.e., Subject-Predicate-Object. This makes it possible to directly search for articles with precisely such a relationship. In contrast, when only MeSH is used as the basis for annotation, a document would be simply labeled, *independently*, with Subject and Object. This would be searched for by a Boolean combination like “Subject AND Object” (This is not the exact syntax used by PubMed for MeSH term-based queries, but serves to convey the concept). This would not only return documents similar to that returned by MachineProse but also return false positive articles that refer to

Subject and Object, but in independent contexts. The difference will be the most marked in cases where Subject and Object are fairly common terms more likely to occur in the same document by random chance. This kind of false positive retrieval of documents is only likely to worsen with the rapidly growing size of the MEDLINE repository. To be fair, the MeSH system does include subheadings or qualifiers (MeSH Browser) that restrict the scope of Subject or Object to a subarea – but this does not apply to the assertion as a whole.

The UMLS (Humphreys et al. 1998; Lindberg et al. 1993) consists of three Knowledge Sources of which we are interested in two, the Metathesaurus (Schuyler et al. 1993) and the Semantic Network (McCray et al. 2001). The Metathesaurus is a unified collection of many different medical terminologies (the January 2003AA edition includes 875,255 concepts and 2.14 million concept names in over 100 biomedical source vocabularies). The Semantic Network contains 135 semantic types (e.g., Disease or Syndrome, Virus). These semantic types (i.e., broad categories) are organized into a hierarchy of IS-A links and 54 kinds of non-IS-A relationships are used to relate them (e.g., Virus *causes* Disease or Syndrome) to each other. Since every concept in the Metathesaurus is assigned to at least one, but often several, semantic types in the Semantic Network, classifying the discrete concepts in the assertion can be done using the Semantic Network. UMLS does provide implicit support for assertions in the form of relationships like “X IS-PART-OF Y” or “X IS-A Y” but the universe of assertions is limited by the structure of the concept hierarchies and the current number of relational operators.

MachineProse seeks to increase specificity by offering a placeholder hierarchy for assertions.

This is made possible by a combination of two knowledge structures. The MachineProse Ontology (MPO) is an Ontology of scientific assertions based on hierarchical Relationship types (Fig. 2). The second is a knowledge-base of instances of scientific assertions that we call the

MachineProse Trove (MPT) shown in Figs. 1 & 3. The MPO is meant to represent a highly refined view of relationships focused on capturing scientific assertions. This is in contrast to the Semantic Relations of the Unified Medical Language System (UMLS) Semantic Network (Browne et al. 2003) which have only 54 types of fairly coarse resolution, with many important types being absent. For instance, the important concept of “Regulates” (Fig. 2) is not represented as a relation. The term does occur in the MeSH hierarchy, but as the concept of “Social Regulation.” We derive relationship concepts from the UMLS Network to MPO wherever possible, but MPO is more comprehensive and optimized to represent assertions. The initial set of 54 semantic relations has been expanded to over 100 (Table 2), constituting different kinds of relationships found in assertions.

Additional inferencing over relationship hierarchy in MachineProse. One of the advantages offered by an Ontology is the ability to carry out reasoning over instances. Ontologies like the Gene Ontology (GO) (Ashburner et al. 2000) and MeSH topics have a hierarchy to describe terms in the Ontology, but both support limited types of relationships. GO supports IS_A or PART_OF relationships while MeSH topics only support IS_A. For example, a query with the MeSH term “Hypersensitivity” also implies searching for articles with its synonyms “Allergy” and “Allergic reaction.” In addition, since “Immune Complex Disease” IS_A form of “Hypersensitivity” based on the MeSH hierarchy, the automatic inference made is that documents containing “Immune Complex Disease” should also be searched for. MachineProse takes this a step further by offering a unique kind of reasoning that exploits hierarchies of relationships *per se*. For instance, since “Regulates” is a superclass of both “Stimulates” and “Inhibits,” a search for an assertion involving regulation, i.e., “A regulates B”, will also search for documents that have the assertions “A stimulates B” and “A inhibits B.” This is equivalent to

a more specific form of the general assertion “Disease is Associated with Protein.” Just like Chronic auto-immune urticaria and anti-thyroid antibody are specific subclasses of Disease and Protein, so is the relationship *Caused* a more specific form of *Associated*. In this respect, MachineProse is similar to the Galen Ontology based on GRAIL (Rector et al. 1997) which offers a rich variety of relationships like “actsOn” and “hasLocation.” However, Galen has primarily focused on the representation of medical anatomy and procedures, while MachineProse is customized for scientific assertions reported by research groups. The UMLS semantic network keeps track of semantic inverses of relations, e.g., IS-PART-OF being the inverse of CONTAINS. This can be used for some degree of inferencing for assertions that are based on these relational operators.

Representation of complex assertions. Many, but not all, assertions can be cast in the simple form of Subject-Predicate-Object (rounded rectangle in Fig. 1). To fully express more involved assertions, we propose the following preliminary version of grammar for an assertion, expressed in Backus-Naur format:

Assertion ::= Entity Relationship Entity [AssertionQualifier]*

Entity ::= SimpleEntity [QuantityQualifier] [EntityQualifier] [(EntityCombinationConstraint | EntityConnector) SimpleEntity]*

SimpleEntity ::= "MeSH Term" | "GO Term" etc.

QuantityQualifier ::= "High dose" | "Low dose" | etc.

EntityQualifier ::= "Near-fatal" | "Sporadic" etc.

EntityCombinationConstraint ::= "Not"

EntityConnector ::= "And"

Relationship ::= "Relation" [BeliefQualifier]*

Relation ::= "Causes" | "Cures" | "Treats" | "Stimulates" | etc.

BeliefQualifier ::= "Maybe" | "Very" | "Strong" | etc.

AssertionQualifier ::= PlaceQualifier | TimeQualifier | StageQualifier | GroupQualifier | ConditionQualifier

Thus, an assertion like “Menstruation may be associated with near-fatal episodes of Asthma” (Martinez-Moragon et al. 2004) may be recast as “Menstruation is associated [BeliefQualifier: Maybe] with Asthma [EntityQualifier: near-fatal episode].” This has some resemblance to MeSH qualifiers (MeSH Browser), but improves on it in two different ways. Firstly, it is much more flexible expressive. Secondly, it can be applied to just the entities (like the qualifiers) or to the assertion as a whole (novel feature). To maintain machine-readability of such a representation, the best choice would be to have a numeric representation in standard units, e.g. dose of a drug in milligrams for the quantity qualifier. Since this is not possible for all qualifiers, an alternative would be to adopt a standard ordinal vocabulary for the terminal tokens whenever possible, e.g., for the quantity and belief qualifiers. When strict ordering is not possible, a partially ordered vocabulary may be used. For nominal values, a controlled vocabulary with synonyms could be used. For both ordinal and nominal values, one should ideally find a way to anchor each ordinal value to a numeric range or value (e.g. Grade 1 Fever being in the range of 38 to 39 degrees Celsius) for inter-annotator (curator or community) consistency. Anchoring is not always possible, but when applied appropriately, will improve the reliability of inferencing. Ideally, growth of such a vocabulary should be curated. As a compromise, new values could be indexed as text words with curation incorporating them into standard terms as the frequency of their use increases. Thus, standardization would be user-driven, but not arbitrary. In terms of implementation, some of its features, e.g., BeliefQualifier, could be adopted by being stored within the additional Relationship Attribute column of the UMLS’s MRREL table (Liu et al. 2002). However, as currently proposed, the machine-readability of complex assertions represented in MachineProse, while superior to existing strategies, is limited in scope. At the

least, some degree of free-text indexing in the space of documented assertions might still be required.

MachineProse captures emerging research findings. A key difference between MPT and existing Ontologies is that the latter are typically based on established domain knowledge (Fig. 3). Typically, domain knowledge lags behind the latest research and evolves gradually over time. For example, updates are indeed made to the MeSH vocabulary (Browne et al. 2003) at periodic intervals. These updates are primarily based on textbooks and encyclopedias. In contrast, the MPT, by definition, is highly dynamic, and meant to capture findings as soon as they are reported in research literature, and incorporate them into the MPT. Thus, MachineProse tracks nascent theories which represent the cutting-edge of scientific enquiry; some of which will be subsequently overturned or found to be trivial and fade into obscurity, while others will become part of tried-and-true domain knowledge.

C. MachineProse Framework

Fig. 1 shows the role of the different modules in the MachineProse architecture. The large oval on the top represents the various Ontologies that serve as the namespaces for the entities constituting an assertion. For example, the Gene Ontology can be used for gene products. The relationship type for an assertion is drawn from the MPO. The large rectangle towards the bottom represents the MPT. The small rounded rectangle represents an assertion that may be used for annotation and semantic indexing of abstracts in the database of documents. The lines that connect entities and relationship types represent ontological dependencies between instances of assertions in the MPT. The user interface shown on top interacts mainly with the MPT but also interacts with the other modules.

D. Populating MachineProse Trove

The MPT is created bottom up and incrementally. A hierarchy of assertion types is created by progressive abstractions. This is machine-readable since all the terms come from a controlled vocabulary in a conceptual structure of assertions. Therefore, the heart of MPT consists of a dynamic knowledge structure that utilizes the existing framework of UMLS, MeSH descriptors, qualifiers and entry terms (synonyms), but adds the dimension of assertions. In particular, the emphasis is on using the rich hierarchy of relational clauses specified by MPO, with vocabulary for entities drawn from existing (parent) controlled terminologies and Ontologies (UMLS, MeSH and GO) as well as refinements made by us. The guiding principle for adding an assertion to the MPT is to use the most specific terminologies that are applicable. Given an assertion <S-R-O> involving Subject S and Object O related via Predicate R as stated in a Publication P that is either affirmed (y), negated (n) or deemed inconclusive (i), one of the following cases will apply:

- 1) If an assertion X in MPT is a perfect match for the assertion, then update X with a pointer to P.
- 2) If there is no perfect match but the entities S and O and the relation R can be expressed explicitly using the terminologies of the existing ontologies, then create a new assertion.
- 3) If there is no explicit matching terminology for the entities S or O or no match for the relation R but they are subclasses of existing vocabulary, the new terminologies or relation are added to the MPO and a corresponding assertion created. The newly created terminologies or relation are linked to the respective immediate superclass in the existing ontologies. User involvement is required in selecting the most relevant terminology or relation type. In effect, this reflects simultaneous refinement of both existing ontologies and the MPO to keep up with the appearance of new knowledge.

4) The assertion cannot be expressed in concepts and relations of the ontologies. In this rare event (given here for completeness), it is placed in an orphan group of assertions pending major revision of the ontologies.

E. Searching the MachineProse Trove

The framework makes it possible to support the following kinds of queries:

- 1) A user may specify a putative assertion as a query. Matching assertions are retrieved, together with the links to the respective articles that support, refute or are inconclusive with respect to the assertion. This makes it easier to retrieve groups of articles relevant to a clinical issue requiring an evidence-based answer, or a specific biological question.
- 2) This design facilitates data mining. For instance, the following kinds of queries can be executed:
 - a. Which are the commonest assertions made in the past year (or month etc.)?
 - b. Which are the most controversial assertions, i.e., have a large number of papers both supporting and refuting an assertion?
 - c. Which assertions are recent but rare? A subset of these might suggest emerging directions for research.

Further, given a query <S-R-O> where at least one of the entities is specified, two dimensions of inferencing are supported:

- 1) Searching assertions through relational inference: For instance, given an assertion query <S-Regulates-O>, 'Regulates' is a more general relation compared to 'Stimulates' according to the MPO. Thus, the search will also include assertions like <S-Stimulates-O>. Similarly, inference from more specific to more general properties is also possible. If a search for the

assertion X does not find a match or yields insufficient results then one can traverse upward in the MP Ontology to find more general relations satisfying the requested assertion.

- 2) Searching assertions through terminological inference: Starting with an assertion $\langle S-R-O \rangle$, one can find more specific assertions defined with refined versions of the terms S or O .

Alternately, one could find generalized versions of the terms as well.

Considering Subject, Object and Predicate together, 3 kinds of matches are possible between a query assertion $\langle S-R-O \rangle$ and entries in MPT – exact, inferred, and partial. An exact match denotes a perfect match with the complete assertion, e.g., $\langle S-R-O \rangle$ in Fig. 4. Inferred matches are of 2 types. They are either descendant matches or immediate parents. A descendant match refers to all assertions that satisfy $\langle dS|S| - dR|R| - dO|O| \rangle$, where the prefix ‘d’ denotes descendant (recursive child). Examples of these are $\langle dS-R-dO \rangle$ or $\langle dS-dR-dO \rangle$. It is logical to consider all descendants down to the leaf/terminal level as matches. For example, if the query is phrased as ‘Adrenal_Cortex_Hormone associated_with Lung_disease,’ it makes sense to consider assertions involving both glucocorticoids and mineralocorticoids that prevent or improve (and other recursive subclasses of ‘associated_with’, Fig. 2) different kinds of lung diseases, including but not limited to Asthma. Another kind of descendant is the conjunction of either entity (Subject or Object) with a second one, e.g., ‘Adrenal_Cortex_Hormone AND Azathioprine associated_with Lung_disease.’ We consider this to be a descendant by multiple inheritance. In terms of the search space of assertions, this may be logically considered to be a subspace of ‘Adrenal_Cortex_Hormone associated_with Lung_disease.’ Note that in terms of implementation, this necessarily creates non-unique paths to each assertion (consequent to being a Directed Acyclic Graph structure rather than a Tree), with the potential for redundancy and inefficiency in mapping and inferencing. Immediate parents are assertions that satisfy $\langle pS|S-$

pR|R-pO|O>, e.g., <pS-R-O>. We restrict to the match to immediate parent because the retrieved assertions from higher superclasses will be too general to be considered a match. For example, for the same query as above, ‘Hormones associated_with Lung_disease’ would be returned but if we navigated higher, we would end up at the root, with essentially any MeSH term being considered a match. A partial match involves either S or O from the original query, and substitutes wild cards for the other components of the triplet. In this case, the specificity of the search can be controlled by an ontology lookup function that determines the semantic distance of two terms in existing ontologies like MPO, UMLS and GO. This is proportional to the number of intervening terms. All potential matches for a given assertion <S-R-O> are returned, ranked by the relative importance of Subject, Object and Predicate, parameterized by weights.

IV. RESULTS: PROTOTYPE AND CASE STUDY

A. MachineProse Prototype

To establish proof of concept for the proposed model for representing the results of scientific research, we have implemented a prototype that demonstrates a subset of the features described under Design. The MP prototype has been implemented using Java Servlets and Java Server Pages (JSP). The MPO has been implemented in the Web Ontology Language (OWL) using the Protégé editor (Noy et al. 2003). For the purpose of modeling and sharing scientific assertions over the Web, we found OWL/Protégé to be the best option for a good trade off between expressivity and traceability of the precise form of assertions. This is due to the fact that OWL is a Web ontology language based on current semantic web standards that is more expressive than XML, RDF and RDF Schema. Protégé is a widely accepted tool in the medical informatics community used for creating and editing ontologies.

The MPT contains instances of the scientific assertions as <subject, predicate, object>. The MPT knowledge base interface is used for the addition of assertions to MPT. The interface API offers a higher level of abstraction over the Protégé API and the Jena API. It uses the Resource Description Framework

Schema (RDFS) as the backend and the Protégé OWL plugin to provide support for both RDF and OWL ontologies. Jena is used for querying and inferencing on the MPT. The MPT interface includes methods to create classes and instantiate them, link them into assertions and attach instances of publications to them. It also allows recursive listing of subclasses. The ontology handler communicates with external ontologies via the UMLS. It connects to the UMLS API to invoke the lookup service for determining the semantic distance between two concepts, and for retrieving synonym sets of concepts. Pointers to relevant publications are stored and the PubMed MEDLINE server invoked on demand. Due to the lack of maturity in Jena/RDF inference tools, performance may be compromised for large scale applications. To resolve this issue, we are working towards developing a new query/inference model based on combining relational databases and ontological solutions.

The MPT can be accessed by both graphical as well as text-based interfaces. The graphical browser has been designed and implemented using the TGViz tab in Protégé. An HTTP connection to the prototype Web Services is established to access the MPT. This is used both for browsing through the knowledge structure of MPT as well as to view the results of a query. This allows the user to either navigate up towards higher abstraction or down to higher refinement. For instance, a user could start with an open-ended query to get to a neighborhood and browse around it to get a more specific answer. Fig. 5 shows the MPT text-based query interface. The *assertion match* lists exact matches and inferred matches based on subclasses in the terminology and relation hierarchies. For example, results of the query *<azathioprine, effective, chronic asthma>* include one publication for the exact match, as well publications for the immediate parent *<azathioprine, effective, asthma>* and descendants like *<azathioprine AND delagil AND heparin, effective, chronic asthma>*. The *partial match* (not shown) lists assertions which share only a subset of the triplet. For each assertion, hyperlinks to relevant publications indicate whether it is upheld, refuted or considered inconclusive. The screenshot (Fig. 5) shows the

relevant publications for the assertion `<azathioprine, effective, asthma>`. We are collecting feedback from clinicians and researchers for the most useful way to display the results for a full-scale implementation. For instance, it would be helpful to provide the assertion together with publication titles on a single screen, along with tallied counts for ‘P,’ ‘N’ and ‘I.’

B. Case Study

We searched the Cochrane database (Grimshaw 2004) for systematic review articles on Asthma. We found 87 abstracts, each of which addressed a specific question dealing with clinical practice. A total of 114 assertions were derived from these abstracts. This was based on corroboration between the first two authors of the present paper. These were converted into MachineProse syntax and loaded into the MPT prototype. These assertions are labeled as authoritative, to indicate that each assertion represents the result of the meta-analysis of several papers, and therefore is likely to be more reliable than that reported by a single study.

The MachineProse hierarchy (Fig. 6) has been constructed by first mapping the assertion entities in the MPT to MeSH. The mapping has been done in a semi-automatic manner by selecting an appropriate class from the available classes. In fig. 6, “.” represents “AND” (e.g., “Gold.corticosteroid” means “Gold and Corticosteroid”) and “*” is used to describe the context information of entities (e.g., “Inhaled*corticosteroid” indicates that the inhaled mode of delivery is the context for the drug “corticosteroid”). In addition to being an expressive notation, “AND” may also be implemented as a logical “AND.” The asterisk implies filtering searches in the knowledge space tentatively described in BNF in the MachineProse model section. Where no class was found in MeSH, we searched for the existence of potential superclasses and added the term to the most appropriate place. There are 231 classes in the MP hierarchy composed of 184 from MeSH and 47 newly introduced classes (Table 3). The latter is made up of 10 newly

introduced classes, 17 compound term classes (e.g., Gold.corticosteroid) composed of more than one term and 20 classes used to express the context of the assertion (e.g., oral, low dose, long acting). The top five MeSH categories used for mapping the assertions include *Hormones, hormone Substitutes, and Hormone Antagonists* (35), *Chemical Actions and Users* (20), *Therapeutics* (10), *Immunologic and Biological Factors* (9), and *Polycyclic Compounds* (8). The MeSH ID is denoted with symbols “[“ and “]”. Thirty five assertions belong to the *Hormones, hormone Substitutes, and Hormone Antagonists* category - of which 24 are positive, 4 are negative and 7 inconclusive. Considering all 114 assertions, about 50% are reported as positive, 15% as negative and 33% as inconclusive.

Fig. 2 shows the hierarchical nature of the MachineProse Ontology (MPO) that is constructed based on extending the set of UMLS semantic relations. In fig. 2, relationship types with 54 ID starting with “T” are from the original set of 54 in UMLS and those with ID starting with “MP” are some of the newly introduced 51 types to represent scientific assertions. For example, the relational concept of “regulates” has been refined into 10 sub-types to denote whether the effect is due to decreased/increased synthesis or degradation, or by modulation of activity. The entities related by an assertion involving one of these types of “regulates” could potentially be a protein, gene, transcript, or chemical compound. Table 2 shows that most of the new relation types were added under ‘Associated With.’ This is because this is the broad type of relational clause that constitutes the hypothesis and conclusion of many biomedical, and especially medical papers.

Fig. 7 shows a network view (Batagelj et al. 2003b) of the MPT, which is composed of a number of assertions. Consider the two Assertions A14 <Dietary_intervention, *improves*, Asthma.atopic_disease> and A13 <Low*salt, *effective*, Asthma>. The subject *Dietary intervention* is mapped to the MeSH term *Caloric Restriction [E02.642.249.200]* and the subject

*low*salt* is mapped to the MeSH term *Diet, Sodium-Restricted* [E02.642.249.510]. They are a subclass of *Diet Therapy* [E02.642.249]. The predicate *improve* is a subclass of the predicate *effective_in* in the MPO. The object *asthma.atopic_disease* is a subclass of *asthma* in the MP hierarchy. Thus, the three entities in the assertions <Subject, Predicate, Object> are hierarchically related to each other through the MPO and MP hierarchies. It is of interest to note that there are relationships between assertions as well (considering the entire assertion as a unit). For example (Fig. 7), the assertions A97 <Speleotherapy, *Effective*, Asthma>-2001-I and A98 <Speleotherapy, *Effective*, Asthma>-2000-I are related as a time series in 2000 and 2001. “I” indicates that the assertion is an inconclusive assertion, i.e, it is not clear whether speleotherapy is effective or not in asthma.

V. Discussion

A. Advantages of MachineProse

Consider the question “Is azathioprine effective in asthma?” This is an important issue that has been the subject of a Cochrane review (Dean et al. 2004). On trying a PubMed query with the words “Azathioprine AND effective AND asthma,” only 3 hits were retrieved. On the other hand, the query “Azathioprine AND Asthma,” retrieved 58 hits. We ignored 22 of these as they did not contain an abstract. Of the remaining 36, only 12 reported assertions relating azathioprine and asthma. A similar analysis for a query about the effect of breast milk in lowering the probability of developing asthma is shown in Table 4 (In this case, adding the word ‘Prevents’ or ‘Protects’ resulted in only 2 and 4 hits respectively). Precision is low in both cases. By design, MachineProse can explicitly handle the query <Azathioprine, *Effective*, Asthma> and return highly specific results. Essentially, at a glance, the assertions made by different publications are obvious. In this particular case, the domain experts on our team had to read each abstract to

extract this information. In contrast, this information would be readily available if articles had been annotated in MP syntax. Thus, MPT makes possible a high degree of precision and recall for semantic queries, returns assertions *per se* (and not just pointers to literature), and states whether publications are in favor of, against, or inconclusive with respect to each assertion. We next did a rough assessment of the proportion of biomedical publications that can be summarized by MachineProse. We queried PubMed for all abstracts by authors with the last name “Smith” and read the first 200 (Table 5). Of these, 108 were centered around specific assertions. Of the remainder, 27 had no abstracts, while the rest could not be readily expressed in the form of the assertions. This corresponds to a coverage of 62% - roughly 3 out of every 5 publications. This represents a large proportion of research publications. Articles that could not be cast in the form of an MP assertion were either reports on the development of methodology, descriptive reports of chemical spectra, or broad-based reviews without any specific conclusions. The coverage is considerably higher for articles in the Jan 9th issue of the Journal of Biological Chemistry (Table 5). The few papers that did not have original and central assertions to report were a mini-review, some crystal structure solutions and exploratory studies. It is of interest to note that the current paper *per se* is not amenable to being represented by MachineProse as described here. We hope to address this in future work, as papers on methodology can be seminal as well.

B. Models for adoption of MachineProse

In this paper, we do not seek to prescribe a definitive approach, but only to demonstrate the potential of MP and to engender discussion towards its refinement and adoption. We believe the power of MP lies in its simplicity of representation, and selective focus on the most important aspect of a paper. Ideally, MP will be most effective if adopted universally. In practice, we

envisage several complementary models for adoption of MachineProse in facilitating scientific research. In addition, to deal with the issue of diverse evolving vocabularies, MachineProse will need to maintain local repositories for new terms that are not part of a standard vocabulary (UMLS, MeSH and GO) as enrichment and refinement (synonym or related terms, specific terms, etc) of existing ontologies.

Co-submission of MachineProse assertions with journal article publication. In addition to keyword entry at the time of article submission for publication, authors can be requested to add assertions specific to that article in MachineProse. Alternatively, these could be specified in natural language with the journal staff being responsible for encoding into MachineProse. Encoding by authors would be more accurate as they would be most familiar with the research being reported, whereas journal staff might be better at dealing with complex assertions, and in enforcing uniformity. In order to take advantage of this across multiple journals, search interfaces for the MEDLINE database like PubMed and Ovid might need to be modified to add capabilities similar to the prototype discussed here. A precedent for this is the concomitant submission of clinical trials results published in the Journal of the American Medical Association into a database (Carini & Sim 2003; Sim et al. 2004).

Semantic citation community hubs. It is also possible to have domain-specific knowledge bases where users directly submit scientific assertions, with the restriction that a reference to a valid publication is always included. This does not require that the original article was annotated with MachineProse at the time of publication, and has the advantage of being community-driven without copyright violation – as only references or links need be submitted. The Citeseer (Bollacker et al. 1998) resource for publications in Computer Science is an example of a successful precedent. These knowledge bases could be empowered with APIs for mining

assertions *per se* in finding support for a given hypothesis or suggesting new ones. Authors might find this an attractive way to point others to their own work in a distributed workload system. This would ensure scalability and make widespread adoption of such community hubs likely. To handle the issue of standardization as new terms and concepts appear, we will need to ensure a good mapping scheme between terms (analogous to the MeSH browser) that are already present and the terms in a new assertion being added. An approach suggested for extending the GO has been to derive terms from orthogonal vocabularies (Ashburner et al. 2000), but this is not flexible enough for the scope of MP. A practical solution would be to allow users to make extensions to existing vocabulary – but have these flagged for subsequent approval by dedicated curators. Orphan networks (i.e., disconnected) could also be integrated into larger networks by periodic curation. Finally, to bridge multiple Citation hubs for inter-disciplinary analysis, one may have to develop and use an Assertion MetaThesaurus analogous to the UMLS MetaThesaurus, in addition to cross-mapping the entities.

Reverse engineering for older literature. MachineProse is not a substitute for text mining approaches, but a complementary one focused on scientific assertions. It can serve as a standard form to represent the result of text-mining efforts. Approaches similar to that used by Textpresso (Muller et al. 2004) can be used to glean assertions from literature. Several other precedents are based on manual curation (Bader et al. 2003; Chen et al. 1997; Joshi-Tope et al. 2005). We are developing methods to do this. In addition, machine learning can be helpful in partial automation of converting assertions in natural language into MachineProse syntax. An important distinction to bear in mind that many Text-Mining approaches are inherently probabilistic, while source encoding with MP represents a deterministic solution to knowledge representation.

C. MachineProse in Context: Challenges and Possible Solutions

The successful adoption of community initiated efforts (like the first and second models discussed above) will in part depend on a friendly and intuitive interface that will obviate the need to learn MPT syntax and conventions. Even partially correct mapping will decrease the workload of the curators. Another issue to consider is the granularity and density of assertions per paper. Should one submit all assertions reported in a paper? Ideally, yes. Pragmatically, we believe it is best to restrict the reporting to the assertion(s) that justify the publication of the paper for the following reasons: i) This is most likely to be the finding of enduring importance – even though in some cases minor findings that are omitted from the abstract might well turn out to be the most important, ii) Authors are more likely to comply with contributing to the MPT – we don't want the effort of submission to be greater than the actual research, and iii) This will do the best job of highlighting nascent findings; reporting secondary, and perhaps well known assertions might result in lowering the semantic signal to noise ratio for new discoveries. Once the important assertions have been selected, it is perhaps best to encode them in the greatest detail (finest granularity) possible for maximum benefit to the research community, for subtle differences could be responsible for large differences.

A current limitation of our prototype is the fact that only literal matches for relations are used while searching the MPT. Ideally, we'd like to have a list of synonyms and a canonical form for each concept in the MPO. For example, “induces expression of” should be mapped to “increases expression of.” Also, we need to deal with different definitions of the same entity term in different controlled vocabularies. We currently resolve misambiguity among GO, MeSH, and UMLS terms by always consulting GO first, and then MeSH and UMLS. However, this needs more research in dealing with the general case when multiple Ontologies are used. Another challenge is dealing with terms that do not occur in any of the Ontologies. A related problem is

that of deciding exactly where a new assertion should be added into an existing MPT. Allowing curators to do this will have higher consistency, while having authors do this is more scalable. A pragmatic solution is to have authors do this primarily, but with the provision of flagging assertions that merit attention by curators.

At the present time, keywords are not really verified by the reviewer or publisher with rigor and inaccurate assertions may slip in unnoticed as well. It is plausible that individuals may attempt to propagate subjective prejudices under the guise of science. However, this may turn out to be a non-issue because MachineProse will only accelerate the location of papers that make baseless assertions, and unscrupulous authors will end up casting themselves in a negative light. A simple scheme where end-users could register their disagreement with the basis of an assertion might suffice. Optionally, curators could be alerted to assertions that have generated a large number of objections. Thus, there will be no need for onerous censorship. Ideally, authors would submit their assertions at the time of review, not at the time of acceptance. Thus, in the best case, both pre-publishing and post-publication checks would ensure the validity of assertions with respect to individual papers. For full-scale implementations, the issue of knowledge-base integrity with respect to distributed input will also need to be addressed. MachineProse might be perceived as trivializing scientific knowledge in summarized entire papers in a few assertions. We argue otherwise. Firstly, this is not meant to replace the publication of full-length papers, only to augment them with assertions of their findings. Secondly, we feel assertions represent the best way to summarize the broad impact of research and are often the *raison d'être* for a publication. The aim is not to capture 100% of the knowledge in papers. Nor should one expect to be able to annotate every kind of paper. However, even in using only the simplest of formats, we believe it is possible to provide succinct snapshots of a significant proportion of the literature.

VI. CONCLUSION

We have presented and given a detailed description of MachineProse, a framework constituted from current Ontologies, controlled vocabularies and a new taxonomy/hierarchy of relationship types. It is centered on the idea of using scientific assertions as semantic knowledge constructs for reporting, indexing, and exchanging research findings. This can eliminate a lot of redundant effort invested in re-extracting facts from articles in prose and to transform the way in which we communicate the results of scientific endeavor. Additionally, scientific assertions can serve as the basis for the inline (concomitant) construction of knowledge bases that can be mined for evaluating hypotheses and suggesting new ones for experimentation. Finally, this can serve as a standard format for results from text-mining that can be cast in the form of assertions – for example, prediction of molecular interactions as “Molecule x interacts_with Molecule y.” This would facilitate the development of an integrated knowledge-base based on several algorithms and data sources.

ACKNOWLEDGEMENTS

We would like to thank Dr. Chitra Dinakar for suggesting some of the queries to evaluate, and Sunil Wagh for carrying out some of the literature searches.

Table 1. Assertions in MachineProse

PubMed ID	Type of Assertion	Assertion	Assertion in NL
15106191	Positive	<Anti-leukotriene_agents.corticosteroids, effective, chronic*asthma>	“The addition of licensed doses of anti-leukotrienes to add-on therapy to inhaled glucocorticoids brings modest improvement in lung function.”
14583955	Negative	<Helium.Oxygen,effective, acute*asthma>	At this time, heliox treatment does not have a role to play in the initial treatment of patients with acute asthma.
14973944	Inconclusive	<Acupuncture, effective, chronic*asthma>	“There is not enough evidence to make recommendations about the value of acupuncture in asthma treatment.”

Table 2. Expansion of the UMLS Relation types in creation of the MPO

	UMLS Semantic Network Relations	Newly Introduced Relations
Functionally_associated/Related_with [T166]	20	41
spatially_related_to [T189]	5	1
conceptually_related_to [T158]	13	5
temporally_related_to [T136]	3	1
physically_related_to [T132]	9	2
isa [T186]	1	0
mpo_relations [MP1]	0	1

Table 3. Distribution of assertions and conclusions regarding them with respect to the MeSH hierarchy

Assertion Type	Positive	Negative	Inconclusive	#Assertion
Hormones, Hormone Substitutes, and Hormone Antagonists [D06]	24	4	7	35
Chemical Actions and Uses [D27]	11	5	4	20
Therapeutics [E02]	0	0	10	10
Immunologic and Biological Factors [D24]	4	1	4	9
Polycyclic Compounds [D04]	7	0	1	8
Heterocyclic Compounds [D03]	5	1	1	7
Organic Chemicals [D02]	2	1	1	4
Environment and Public Health [G03]	0	0	4	4
Inorganic Chemicals [D01]	1	1	1	3
Health Services Administration [N04]	2	0	1	3
Health Care Facilities, Manpower, and Services [N02]	1	1	0	2
Human Activities [I03]	1	1	0	2
Behavioral Disciplines and Activities [F04]	0	0	2	2
Animals [B01]	0	1	0	1
Digestive System Diseases [C06]	0	0	1	1
Lipids [D10]	0	1	0	1
Growth Substances, Pigments, and Vitamins [D11]	0	0	1	1
Equipment and Supplies [E07]	0	0	1	1
Total	58	17	38	114

Table 4. Relevancy of documents retrieved by PubMed

PubMed Query	Total	Excluded	Relevant	Precision
Azathioprine AND asthma	58	22	12	33%
Breast milk AND asthma	60	10	21	42%

Table 5. Proportion of papers that can be represented by MP

Abstracts	Total	Excluded	With assertions	Coverage
Sample of the Cochrane review articles searched by “Asthma”	87	0	87 (114 assertions)	100%
Sample of articles by ‘Smith’	200	27	108	62%
Journal of Biological Chemistry (2005)	99	0	89	90%

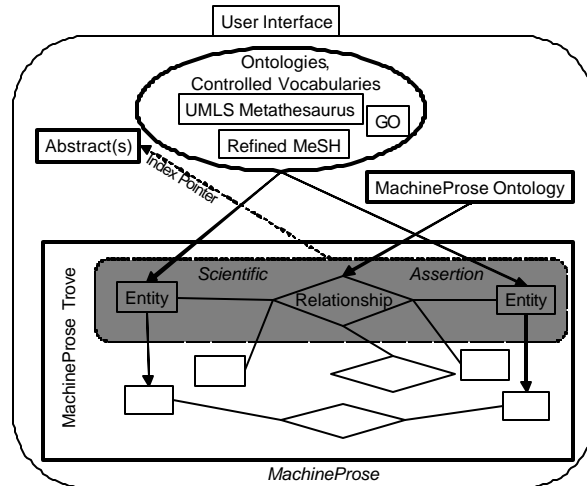


Figure 1. Overview of the MachineProse framework

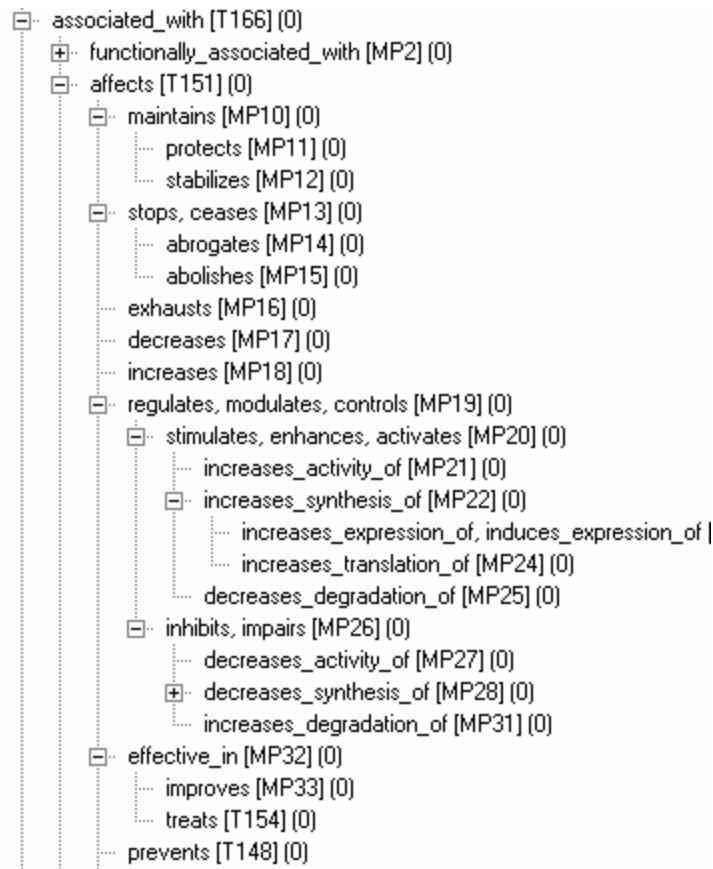


Figure 2. Hierarchical view of the Relationship Types of the MachineProse Ontology

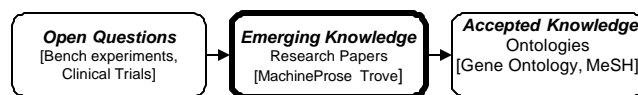


Figure 3. Ideal role of the MachineProse

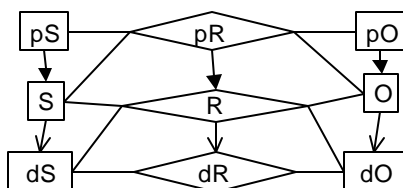


Figure 4. Semantic matches in assertion space

MachineProse Trove Query Interface

Term 1:

Relation:

Term 2:

azathioprine effective asthma

Pubmed ID	Publication Support
7975810	P
24422	P
14974011	I
8420024	F
4023940	P
15227496	P
12086374	P
1470705	F
11109414	P
8404082	P
2083410	P

Search Results

Assertion Matches

- [- azathioprine AND adrenal cortex hormones](#)
- [-- azathioprine AND adrenal cortex hormone](#)
- [--- azathioprine AND adrenal cortex hormor](#)
- [---- azathioprine AND adrenal cortex horm](#)
- [----- azathioprine AND adrenal cortex horn](#)
- [- azathioprine AND delagil AND heparin MPT](#)
- [- azathioprine AND delagil AND heparin AS](#)
- [- azathioprine AND delagil AND heparin F](#)
- [- azathioprine AND delagil AND heparin](#)
- [- azathioprine AND delagil AND heparin](#)
- [- azathioprine AND delagil AND heparin](#)
- [- azathioprine MPT RELATIONS asthma \(1](#)
- [-- azathioprine ASSOCIATED WITH asthr](#)
- [--- azathioprine FUNCTIONALLY RELAT](#)
- [---- azathioprine AFFECTS asthma \(11](#)
- [----- azathioprine EFFECTIVE asthma \(1](#)
- [----- azathioprine EFFECTIVE chronic asthma \(1](#)

Partial Matches

Publication Support Key	
P	Positive
N	Negative
I	Inconclusive

Figure 5. Assertion matches for the query <Azathioprine, effective, Chronic asthma> retrieved by the MPT prototype.

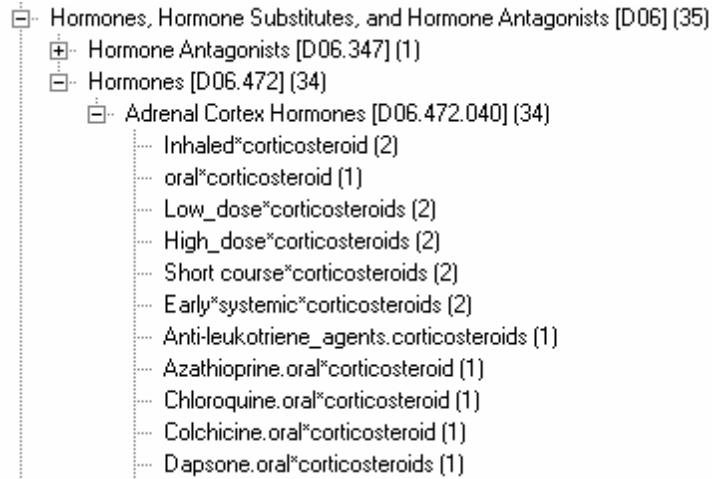


Figure 6. Mapping of terms in assertions to the MeSH hierarchy

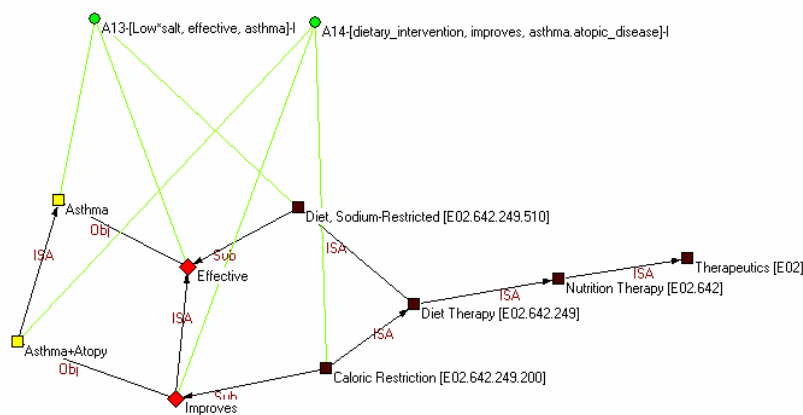
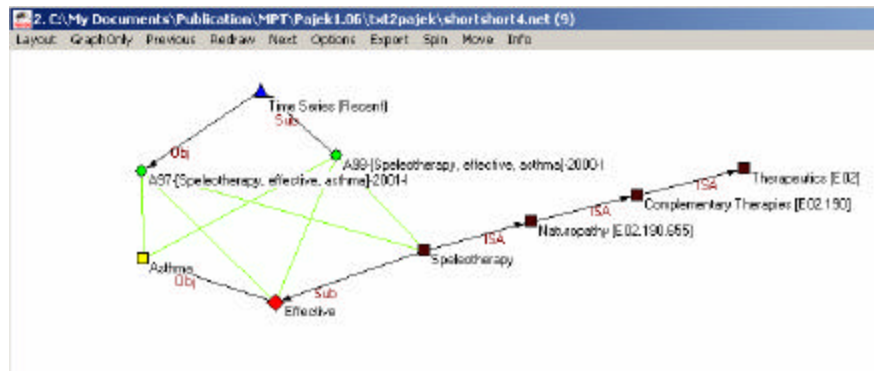


Figure 7. Network view of the MachineProse Trove

REFERENCES

- Aronson AR, Bodenreider O, Chang HF et al. The NLM Indexing Initiative. *Proc AMIA Symp* 2000;17-21.
- Aronson AR, Mork JG, Gay CW et al. The NLM Indexing Initiative's Medical Text Indexer. *Medinfo* 2004; 11 (Pt 1):268-72.
- Aronson AR, Rindfleisch TC. Query expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp* 1997:485-9.
- Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; 25 (1):25-9.
- Bachrach CA, Charen T. Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. *Med Inform (Lond)* 1978; 3 (3):237-54.
- Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 2003; 31 (1):248-50.
- Batagelj V, Mrvar A. Analysis and Visualization of Large Networks. In: P Mutzel, editor, translator and editor *Graph Drawing Software*. Berlin: Springer; 2003a; p. 77-103.
- Batagelj V, Mrvar A. Pajek - Analysis and Visualization of Large Networks. In: M Junger; P Mutzel, editors, translator and editor *Graph Drawing Software*. 1st edn. Berlin: Springer; 2003b; p. 77-103.
- Benson DA, Karsch-Mizrachi I, Lipman DJ et al. GenBank: update. *Nucleic Acids Res* 2004; 32 Database issue:D23-6.
- Berman HM, Westbrook J, Feng Z et al. The Protein Data Bank. *Nucleic Acids Res* 2000; 28 (1):235-42.
- Bernstein P, Cohen B, MacCallum C et al. PLoS Biology-We're Open. *PLoS Biol* 2003; 1 (1):E34.
- Boeckmann B, Bairoch A, Apweiler R et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003; 31 (1):365-70.
- Bollacker KD, Lawrence S, Lee Giles C. CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. In *2nd International ACM Conference on Autonomous Agents*: ACM Press; 1998. 116-23 p.
- Browne AC, Divita G, Aronson AR et al. UMLS language and vocabulary tools. *AMIA Annu Symp Proc* 2003:798.
- Carini S, Sim I. SysBank: a knowledge base for systematic reviews of randomized clinical trials. *AMIA Annu Symp Proc* 2003:804.
- Chen RO, Felciano R, Altman RB. RIBOWEB: linking structural computations to a knowledge base of published experimental data. *Proc Int Conf Intell Syst Mol Biol* 1997; 5:84-7.
- Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005; 6 (1):57-71.
- Dean T, Dewey A, Bara A et al. Azathioprine as an oral corticosteroid sparing agent for asthma. *Cochrane Database Syst Rev* 2004; (1):CD003270.
- Divoli A, Attwood TK. BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics* 2005; 21 (9):2138-9.
- Eisen MB, Brown PO, Varmus HE. Public-access group supports PubMed Central. *Nature* 2002; 419 (6903):111.

- Fiszman M, Rindflesch TC, Kilicoglu H. Integrating a hypernymic proposition interpreter into a semantic processor for biomedical texts. *AMIA Annu Symp Proc* 2003:239-43.
- Grimshaw J. So what has the Cochrane Collaboration ever done for us? A report card on the first 10 years. *Cmaj* 2004; 171 (7):747-9.
- Humphreys BL, Lindberg DA, Schoolman HM et al. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc* 1998; 5 (1):1-11.
- Joshi-Tope G, Gillespie M, Vastrik I et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005; 33 (Database issue):D428-32.
- Keseler IM, Collado-Vides J, Gama-Castro S et al. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res* 2005; 33 (Database issue):D334-7.
- Krauthammer M, Kra P, Iossifov I et al. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* 2002; 18 Suppl 1:S249-57.
- Krieger CJ, Zhang P, Mueller LA et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 2004; 32 (Database issue):D438-42.
- Libbus B, Rindflesch TC. NLP-based information extraction for managing the molecular biology literature. *Proc AMIA Symp* 2002:445-9.
- Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993; 32 (4):281-91.
- Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assoc* 2002; 9 (6):621-36.
- Martinez-Moragon E, Plaza V, Serrano J et al. Near-fatal asthma related to menstruation. *J Allergy Clin Immunol* 2004; 113 (2):242-4.
- McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. *Medinfo* 2001; 10 (Pt 1):216-20.
- A MEDLINE Indexing Experiment using Terms suggested by MTI. 2002. <http://ii.nlm.nih.gov/resources/ResultsEvaluationReport.pdf>
- MeSH Browser.
- Mika S, Rost B. Protein names precisely peeled off free text. *Bioinformatics* 2004; 20 Suppl 1:I241-17.
- Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2004; 2 (11):e309.
- Nelson SJ, Schopen M, Savage AG et al. The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation. *Medinfo* 2004; 2004:67-9.
- Noy NF, Crubezy M, Ferguson RW et al. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu Symp Proc* 2003:953.
- Rebholz-Schuhmann D, Kirsch H, Couto F. Facts from text--is text mining ready to deliver? *PLoS Biol* 2005; 3 (2):e65.
- Rector AL, Bechhofer S, Goble CA et al. The GRAIL concept modelling language for medical terminology. *Artif Intell Med* 1997; 9 (2):139-71.
- Revere D, Fuller S, Bugni PF et al. An information extraction and representation system for rapid review of the biomedical literature. *Medinfo* 2004; 11 (Pt 2):788-92.
- Rindflesch TC, Aronson AR. Semantic processing in information retrieval. *Proc Annu Symp Comput Appl Med Care* 1993:611-5.
- Rindflesch TC, Libbus B, Hristovski D et al. Semantic relations asserting the etiology of genetic diseases. *AMIA Annu Symp Proc* 2003:554-8.

Rindflesch TC, Tanabe L, Weinstein JN et al. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput* 2000:517-28.

Rzhetsky A, Iossifov I, Koike T et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* 2004; 37 (1):43-53.

Schuyler PL, Hole WT, Tuttle MS et al. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bull Med Libr Assoc* 1993; 81 (2):217-22.

Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. *J Biomed Inform* 2004; 37 (2):108-19.

Sneiderman CA, Rindflesch TC, Aronson AR. Finding the findings: identification of findings in medical literature using restricted natural language processing. *Proc AMIA Annu Fall Symp* 1996:239-43.

Srinivasan P, Rindflesch T. Exploring text mining from MEDLINE. *Proc AMIA Symp* 2002:722-6.

Zhou G, Zhang J, Su J et al. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* 2004; 20 (7):1178-90.

Zhu AL, Li J, Leong TY. Automated knowledge extraction for decision model construction: a data mining approach. *AMIA Annu Symp Proc* 2003:758-62.

FIGURE LEGENDS

Figure 1. Overview of the MachineProse Framework. Various components of the MachineProse framework as shown. The triplet of Entity (small rectangle) - Relationship (diamond) - Entity represents a Scientific Assertion. The Entity terms are derived from parent ontologies and controlled vocabularies while the Relationship terms are derived from the MachineProse Ontology. Several Scientific Assertions are organized into a hierarchical network to form the MachineProse Trove (MPT). Each assertion in the MPT is used to annotate at least one scientific publication, and serves as a semantic index. The framework is accessed through a user interface.

Figure 2. Hierarchical View of the Relationship Types of the MachineProse Ontology. Some of the Relationship Types that constitute the MP Ontology are shown. These have been extended from the original set of Semantic Relations that are part of the UMLS.

Figure 3. Ideal Role of the MachineProse Trove. The ideal role of the MachineProse Trove is to capture emerging discoveries as soon as they are published. This is in contrast to the role of domain Ontologies that conceptualize well-established facts.

Figure 4. Semantic Matches In Assertion Space. The figure shows how inferencing proceeds from an assertion comprised of Subject (S), Relationship (R) and Object (O). The prefix 'p' denotes an immediate parent while the prefix 'd' denotes a descendant term. The latter includes children as well as recursive children. Implicitly, all descendants are matches for a query that is framed as S-R-O. Additionally, immediate parents are also included as putative matches.

Figure 5. Assertion matches for the query <Azathioprine, effective, Chronic asthma> retrieved by the MPT prototype. The figure shows two panels of the prototype search interface for the

MPT. The retrieved assertions include azathioprine in conjunction with other drugs as well as considered separately. The panel in front shows the PubMed IDs of 11 scientific articles that have evaluated this assertion, the majority of which report that this assertion is true.

Figure 6. Mapping of Terms in Assertions to the Mesh Hierarchy. This shows the number of entities occurring in the assertions, together with their associated qualifiers, mapped to appropriate locations within the MeSH hierarchy.

Figure 7. Network view of the MachineProse Trove. A graphical view of parts of the network of assertions (MPT) has been created using Pajek (Batagelj et al. 2003a). Some of the assertions represented in the figure are “Caloric restriction improves Atopy,” and “Speleootherapy is effective in asthma.”