# Toward a Knowledge Representation Corpus of Historical Events

Robert C. Kahlert
Cycorp Inc.
*rck@cyc.com*

Jennifer Sullivan
Convera Inc.
*jsullivan@convera.com*

November 2005

**Abstract**

Research communities need a large corpus of representative, relevant and interesting problems to evaluate their proposed solutions; unfortunately the KR&R community lacks such a corpus. We therefore propose to construct a large corpus of knowledge representation and reasoning problems, drawing upon readily available historical real-world events for contents, in a highly expressive representation language such as CycL. We discuss some of the properties that the KR&R corpus and the chosen historical events should have to support KR&R research and suggest a specific historical event, the Salem Witch trials, as an appropriate tracer bullet for the construction of this corpus.

## 1 Introduction

Large corpora of information have galvanized the information retrieval and natural language processing communities. Projects such as the TREC evaluations of NIST [TREC] or the PENN Treebank [9] provide the key benchmarks against which innovations are measured. Within the theorem proving community, the TPTP (Thousands of Problems for Theorem Provers) repository, established in 1993 and maintained by the University of Miami, fulfills a similar role. Problem sets for the annual theorem proving competition CASC are drawn from this corpus. In the words of co-maintainer Geoff Sutcliffe [TPTP]:

> The principal motivation for the TPTP is to move the testing and evaluation of ATP [i.e. Automated Theorem Proving] systems from the previously ad hoc situation onto a firm footing. This became necessary, as results being published do not always accurately reflect the capabilities of the ATP system being considered. A common library of problems is necessary for meaningful system evaluations, meaningful system comparisons, repeatability of testing, and the production of statistically significant results. The TPTP is such a library.

Unfortunately, the KR&R community lacks a comparable corpus of problems and solutions to rally to.[1] Issues such as inference or truth maintenance speed, ease of representation or modification, complexity of theory revision or completeness of answers are worked and reported on without reference to a common problem set.

At least part of the problem is that the domain choice of such a corpus is tricky. Preferably, the corpus would use real-world entities and events, whose properties are well documented, so that the corpus can pose realisitic and interesting problems – e.g. temporal and spatial reasoning, causality and argumentation. However, data of current agents and events can be legally difficult to obtain and incur the charge of engaging in profiling. Synthesizing realistic data sets adds to the corpus construction cost.

We therefore propose a knowledge representation corpus that draws upon historical events. The academic historical community already researches and documents these events and their agents in great detail. While some things do change over time, we claim that the types of KR problems one encounters – e.g. script recognition, model revision and contradiction detection – are very similar to the problem types the KR&R community is already pursueing.

The remainder of this paper analyzes the requirements that such a corpus should satisfy from the knowledge representation and reasoning side. A tracer-bullet analysis of a specific historical event will provides a preliminary

[1] The CSR (Common Sense Reasoning) section of the TPTP problem corpus consists of 24 problems of under 50 axioms each, involving water flowing into a kitchen sink and spinning trolleys in a supermarket. In the KRS (Knowledge Representation) domain, 155 of the 175 examples where contributed by Sean Bechhofer and his colleagues [16] and are classification tasks, OWL DL consistency and entailment tasks, all reducible to concept consistency w.r.t. an empty KB via internalisation. These problems cover just a fraction of the interesting problem domains of KR&R.

idea of the categories of KR&R problems that such a corpus can supply to the community. We conclude with some remarks about how to construct such a corpus.

## 2 KR Properties of a History Corpus

We will now look at the properties that a good corpus of historical problems, appropriate for performing knowledge representation and reasoning research, should have.

We propose that such a corpus should consist at least of the following:

- a set of *questions* about events in human history, e.g. "How did Wellington react to the report of Napoleon's death?"
- the *foreground knowledge* for each question needed to understand and attempt to answer the question, e.g. "Wellington fought Napoleon at the Battle of Waterloo.";
- a set of one or more *answers* to each question, complete with one or more sets of justifications (e.g. proof trees) explaining the reasoning steps, rules and grounded facts used in the answer, e.g. "Wellington cried, because he had admired Napoleon as a great general";
- the *background knowledge* that represents the common information shared between all of the questions, answers and justifications, e.g. "During the Napoleonic Era, Great Britain and France jousted for preeminence in Mediterranean Politics".

In short, the corpus will specify the questions, their answers and the knowledge needed to derive them. All of these elements of the corpus will be represented in a formal knowledge representation language amenable to automated reasoning.[2]

Notice that this corpus organization favors the reasoning aspect of KR&R over representational issues; giving a formal specification of the query makes representational choices that reasonable people need not agree with. There are research issues with how to correctly encode a natural language question in a formal representation as well. Therefore, an extended version of the

[2]ResearchCYC [RCyc] uses this approach for its common sense knowledge tests.

corpus should include simplified natural language descriptions (sometimes called "English zero") of the questions, the answers and the facts and rules involved.[3]

## 2.1 Representational Properties

### 2.1.1 Representational Approach

We propose that the knowledge that is represented to be as *declarative* as possible. It is usually feasible to compile declarative knowledge into procedural knowledge while the reverse is more difficult. For example, FOL theorem provers often determine the application order of predicates from the cardinality of the ground literals, while PROLOG-like systems require the order of predicate application to be specified in the knowledge base.

We furthermore propose that the knowledge be represented as *expressively* and as *succintly* as possible, with an emphasis on the ease of authoring the knowledge. Some higher-order logic (HOL) constructs can significantly simplify the representation task [13], while translating readily into less expressive representations[4] which might not share the decidability problems of HOL.

The focus on expressivity extends to representational choices as well. For example, events should probably be represented in a Davidsonian fashion. For such a representation readily translates to a relational ("action predicate" based) representation, where the event is implied. The reverse translation entails the difficult problem of determining whether two "action predicate"-encoded ground facts are describing the same or different events.

By the same token, the representation should not be limited to binary predicates; reducing higher-arity predicates to a binary-only notion is a trivial task for a machine but cumbersome to to do by hand.

### 2.1.2 Choosing a Representation Language

There exist a plethora of knowledge representation languages, and there is little reason to assume that this will change in the near future, despite strong endorsement that languages like Conceptual Graphs [CGStand], KIF [KIF-Stand] or the flavors of OWL [WebOntRef] enjoy.

[3] "English zero" notations are independently useful for knowledge base construction, akin to a representational pseudo-code.

[4] Provided the question itself does not require higher-order constructions to be represented correctly.

We propose that the knowledge representation be as easily *translatable* into other representations as possible. This requirement is easier to fulfill if the chosen representation is as expressive as possible. Choosing a "representational superset" will simplify the authoring of the corpus. All interested parties can then provide translations from that "superset" into their preferred representation and reasoning language.

Since this is a corpus, the goal is for representation and contents to stabilize relatively quickly. This makes the translation to other representations infrequent, allowing for the translation process to involve expensive reasoning. Translations can also flow back into the corpus repository and be leveraged by others who use similar or equivalent representations – as long as the most expressive remains the normative representation language.[5]

### 2.1.3 Choosing an Ontology

With the representational maximization, the choice of an existing ontology becomes less interesting, except for reducing the authoring effort. ResearchCYC provides a nice starter set as an ontology of common-sense vocabulary and comes with some applicable background knowledge for the purposes of the corpus. However, we are not familiar with any ontology today that is sufficiently fleshed out to handle the representational needs of our corpus "out of the box".

## 2.2 Question Properties

The questions should span the gamut of reasoning tasks that the individual representational languages are designed for, including spatial and temporal reasoning, contextualized reasoning, causal reasoning, event or script matching, consistency checking and contradiction finding, hypothetical reasoning and argumentation (i.e. giving pro and con arguments for a particular answer).

Furthermore, the corpus should be partitionable along multiple dimensions, such as the size of the background knowledge needed, the number

[5]CycL is such a sufficiently expressive language that supports both first-order and higher-order logic in its representation and provides a reference implementation with ResearchCYC [RCyc]. Significant parts of ResearchCYC, which is authored in CycL, have been successfully translated into FOL [13], DAML [14] and OWL, or KIF, satisfying our translation requirement. However, any language at least as expressive as CycL will suffice.

of terms involved, the rules of inference available or the expressivity of the language (such as FOL, HOL or DL) required to state the question. By classifying the individual questions as to their requirements, participants can extract those corpus subsets that they are interested in.

## 2.3 Answer Properties

### 2.3.1 Types of Answers

Most questions should have an answer, but not all. For some the right answer should be "unknown" or "not provable" from the knowledge available in the corpus. This foils exhaustive strategies that are of interest for small domains only.

Answers can have certainty associated with them if the reasoning mechanism can usefully employ probabilistic information. However, a limitation of the proposed domain is the fragmentary nature of the historical record, which may make it difficult to establish useful probabilities that would carry over to problems outside the constraints of the historical corpus.

### 2.3.2 Types of Justification

We propose that expected answers to questions be represented in the representation language as well. All answer justifications should bottom out in sentences in the background or the foreground knowledge. Matching the answer proof of an inference with the baseline proof is then a mere check for presence of the expected supporting sentences.

In addition, there may be "implementation knowledge", specific to a class of theorem provers, which capture the rules of inference a specific proof engine implements in a declarative fashion, so that they can be present in the answer proof as well. For example, an inference engine might add an "implementation rule" for the application of *modus tolens* to the proof, despite the fact that the engine implements *modus tolens* procedurally.

Such an approach has the potential of taking into consideration the various proof procedures employed by the theorem provers in the KR community.

# 3 Choosing appropriate Historical Events

## 3.1 Introduction

Historical events are most useful for the purposes of the KR&R community when they pose problems that are similar to the classes of problems already being studying.

Given this assumption, we can identify several properties that candidate historical events should have. Examples for how an event can fulfill these criteria will are given below, when we analyze the proposed "tracer bullet" event.

As mentioned earlier, the historical event (or series of events) should be well researched and its source materials readily available. Much of the historical information should be readily accessible in English, a language shared by most people in the research community, either in translations or editions. Ideally, much of the information should be available at minimal cost or, better yet, accessible via the web — all of which simplifies corpus construction. The event should be a field of active research, so that new findings or new interpretations can be expected to extend the problem sets. The opinions on the event should be diverse, as should be the proposed interpretations of the historical evidence, which would emphasize "point of view" (modal or contextualized) reasoning.

More pragmatically, the chosen event should be reasonably well-known. At the same time, the event should be such that very few of the now living could be offended by the subject matter.

## 3.2 Salem Witchcraft Trials: A Tracer Bullet

One event in the not-too distant historical past that fulfills all of the above requirements are the so-called Salem witchcraft trials.

### 3.2.1 Pragmatic and Documentary Properties

Thanks to Arthur Miller's play *The Crucible* [10], most people familiar with American literature have at least heard of the event. In the United States especially, the events surrounding the Salem witchcraft trials hold the status

of a mystery story that continues to intrigue.[6]

For an event that occurred over 300 years ago, there is an abundance of accessible documentary evidence for the Salem Witchcraft Trials. Scholarly publication of primary sources started in the 1910s [6] [3] and the Great Depression, when a Works Progress Administration program had the trial records [2] transcribed.

In the late 1990s, Benjamin Ray (University of Virgina) began an e-Text project for the primary sources of the Salem Witch trials [EText], which made materials such as court records and transcripts, images and historical maps accessible as text or as "facsimile" images.

### 3.2.2 Scholarly Debate and KR Problems

There are a surprising number of proposals on how to interpret what happened at Salem. We will briefly enumerate some well-known positions and identify types of knowledge representation problems that analyzing these positions would pose.

Of course, all of the interpreations contribute the standard KR&R problems of classification and consistency [16].

Chadwick Hansen [7] reconstructs the Puritan notions of witchcraft and its interaction with Puritan theology and politics. Because witchcraft, both benevolent ("white") and malevolent ("black") was a functional tool, people could feel under attack by something that they utilized. Hansen also argues that the communities' legal and theological requirements for proof of witchcraft changed as the trial progressed and especially points to the Puritan clergy, whose criticism eventually ended the craze.

- *Theory Modeling*: How was the community's belief in witchcraft structured?
- *Scripts*: What were the functional expectations of benevolent and malevolent witchcraft?
- *Theory Revision*: How did the legal requirements of evidence during the proceedings change? Which occurrences caused the respective authorities to modify their opinions?[7]

[6] Googling for "Salem witch" returns 2.3 million hits, which puts the witch story somewhere between Pearl Harbour (2.7 million hits) and Gettysburg (1.95 million hits). `amazon.com`'s returns 255 books for "Salem Witchcraft".

[7] Note that this question of theory revision also pose problems of *Temporal Reasoning*.

Paul Boyer and Stephen Nissenbaum [1] argue that socio-economic competition between mercantile Salem Town and agricultural Salem Village, and between the two most influential families, the Putnams and the Porters, paved the road for conflict. The bone of contention was the village minister Samuel Parris, whose daughter and niece were the first "victims" of witchcraft. The interpretation of Boyer and Nissenbaum would pose the following types of KR&R problems:

- *Social Network Analysis*: How many of the accused belonged to the Porter sphere of influence; how many of the attacked and the accusers belonged to the Putnam sphere?
- *Spatial Reasoning*: How many of the accused lived closer to Salem Town than to Salem Village; how many of the accusers lived closer to Salem Village than to Salem Town?

Carol Karlsen [8] localizes the trials within the social, political and theological role of women in the Puritan Colonies and within the story of witch trials in Colonial America. One problem in Puritan society was the transfer of land possession, which determined the economic status of the new generation: Fear of economic disenfranchisement underpinned the witchcraft accusations. Karlsen claims that the majority of accused "witches" were women who threatened the orderly transfer of land from father to son. This interpretation would pose the following KR&R problems:

- *Inheritance Analysis*: Which accused could have inherited what land from whom?
- *Precedent Analysis*: What documented "intercepted" inheritances occurred in Salem Village during the preceeding years that the girls could have known about?

Mary-Beth Norton [11] describes the events in Salem as influenced by King Philip's and King William's war between the American Indians, their French-Canadian allies and the English colonists. Massacres of the civilian population become the source of the post-traumatic stress disorder that for Norton explains the behavior of the bewitched girls. Norton's interpretation would add the following KR&R problems:

- *Script Matching*: which of the symptoms described for which of the "possessed" girls matches the clinical description of Post-Traumatic stress disorder?

There exists a whole class of reductionistic interpretations which blame a lack of medical knowledge for the outbreaks. Linnda Caporael [4] argues that the rye at Salem Village was affected with the fungus *Claviceps purpura* and explains the experiences of "attacks" as Ergot poisoning (a.k.a. convulsive ergotism). Laurie Carlson [5] proposes an epidemic of tick-borne *encephalitis lethargica* as afflicting the accusers.

These interpretations are relevant because they have been rejected and their short-comings documented. The interpretations contradict then-available historical evidence or are reducible to fallacies.

- *Contradiction Finding*: What pieces of historical evidence undermine these interpretations? What are the argumentative fallacies? What are the relevant counter-examples that refute the interpretations?

A final class of KR&R problems derives from the research as a whole. For example, historical research proceeds itself in a temporal order: Mary-Beth Norton cites Karlsen, who cites Boyer and Nissenbaum, who in turn cite the work of Hansen.

- *Argument Re-Use*: How did the latter authors work with the interpretations offered by their predecessors? Which arguments did they accept, which did they refute, and which did they ignore?

- *Theory Comparison*: What are the facts on which the interpretations agree? What are the facts that interpretation A cites that contradict or refute interpretation B?

- *Truth Maintenance*: Which parts of what interpretations became untenable as research progressed?

# 4 Conclusions and Outlook

We have agreed with Geoff Sutcliffe that research communities need large problem sets to put themselves on a principled footing in comparing their research. Unfortunately, the KR&R community lacks such a corpus.

We have identified the requirements that such a corpus needs to fulfill, both in terms of knowledge representation properties and in terms of real-world contents. Spending time on synthesizing "realistic" data is unnecessary if there is a abundance of real-world historical research to draw upon.

Based on the content requirements, we identified the Salem witch trials as a well-documented and sufficiently accessible historical problem for a pilot KR&R corpus project. We enumerated the many challenging problems that the KR&R community could expect such a Salem Witch Trials corpus to contain.

We are currently investigating representational approaches for some of the problems our analysis has identified and hope to present this aspect of our work in more detail in the final submission.

A word in closing about realizing the construction of such a corpus. As the publication of some of the key documentation in the Salem Witch trials shows, there is both interest and sponsorship in the humanities for contributions to the Salem Witch Trials research. Alternatively, the corpus construction could be part of an effort to revisit some of the thorny problems of large knowledge base construction efforts [12] [RKF], with historians providing the subject matter expertise.

# 5 Appendix

# References


[1] Paul Boyer and Stephen Nissenbaum. *Salem Possessed*. Harvard University Press, 1976.

[2] Paul Boyer and Stephen Nissenbaum, editors. *The Salem Witchcraft Papers (3 vols)*. Da Capo Press, 1976.

[3] George Lincoln Burr. *Narratives of the Witchcraft Cases, 1648-1706*. Charles Scribner's Sons, 1914.

[4] Linnda Caporael. Ergotism: The satan loosed in salem? *Science*, 192, 1976.

[5] Laurie Winn Carlson. *A Fever in Salem*. Ivan R. Dee, 2000.

[6] George Frances Dow, editor. *Records and Files of the Quarterly Courts of Essex County, abstracted and transcribed by Harriet S. Tapley, 9 vols.* Essex Institute, 1911-1975.

[7] Chadwick Hansen. *Witchcraft at Salem.* Braziller, New York, 1969.

[8] Carol Karlsen. *The Devil in the Shape of a Woman: Witchcraft in Colonial America.* W.W. Norton & Co., New York, 1987.

[9] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[10] Arthur Miller. *The Crucible.* 1953.

[11] Mary Beth Norton. *In the Devil's Snare : The Salem Witchcraft Crisis of 1692.* Knopf, 2002.

[12] Kathy Panton, Pierluigi Miraglia, Nancy Salay, Robert C. Kahlert, David Baxter, and Roland Reagan. Knowledge formation and dialogue using the kraken toolset. In *Eighteenth national conference on Artificial intelligence*, pages 900–905, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.

[13] Deepak Ramachandran, Pace Reagan, and Keith Goolsbey. First-orderized researchcyc: Expressivity and efficiency in a common-sense ontology. In Shvaiko et al. [15], pages 33–40.

[14] Stephen Reed and Douglas B. Lenat. Mapping ontologies into cyc. In *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web, July 2002*, Edmonton, Canada, 2002.

[15] Pavel Shvaiko, Jerome Euzenat, Alain Leger, Deborah L. McGuinness, and Holger Wache, editors. *Papers from the AAAI Workshop on Contexts and Ontologies: Theory, Practice and Applications. Pittsburgh, Pennsylvania, July 2005*, Menlo Park, California, 2005. American Association for Artificial Intelligence.

[16] Dmitry Tsarkov, Alexandre Riazanov, Sean Bechhofer, and Ian Horrocks. Using vampire to reason with owl. In *Lecture Notes in Computer Science*, volume 3298, pages 471–485, 2004.

## 5.1 Webpages

Where possible, we cite webpages via the Internet Archive project to ensure accessibility and stability of the document contents.


**CGStand** Conceptual Graphs ISO standard working document
http://www.jfsowa.com/cg/cgstand.htm]
(via http://web.archive.org/web/20041022083436/)

**EText** Salem Witch Trials Documentary Archive and Transcript Project
http://jefferson.village.virginia.edu/salem/home.html, visited November $5^{th}$, 2005

**KIFStand** Knowledge Interchange Format (KIF)
http://logic.stanford.edu/kif/kif.html
(via http://web.archive.org/web/20041030000733/)

**RCyc** Research Cyc
http://research.cyc.com, visited November $5^{th}$, 2005

**RKF** Rapid Knowledge Formation
http://www.rl.af.mil/tech/programs/rkf/, visited November $5^{th}$, 2005

**TPTP** Thousands of Problems for Theorem Provers Repository
http://www.cs.miami.edu/~tptp/, visited November $5^{th}$, 2005

**TREC** Text Retrieval Conference
http://trec.nist.gov/
(via http://web.archive.org/web/20041128090645/)

**WebOntRef** OWL Web Ontology Language Reference
http://www.w3.org/TR/owl-ref/
(via http://web.archive.org/web/20041120090119/)

**WebOntReq** OWL Web Ontology Language Usecases and Requirements
http://www.w3.org/TR/webont-req/
(via http://web.archive.org/web/20041119090335/)