DARPA-BAA-09-03 Machine Reading
Broad Agency Announcement (BAA)

for

Information Processing Techniques Office (IPTO)
Defense Advanced Research Projects Agency
(DARPA)

****

Table of Contents:

Part I: Overview Information 4
Part II: Full Text of Announcement 5
Section I: FUNDING OPPORTUNITY DESCRIPTION 5
Section II: AWARD INFORMATION 42
Section III: ELIGIBILITY INFORMATION 43
A. Eligible Applicants 43
1. Procurement Integrity, Standards of Conduct, Ethical Considerations, and Organizational Conflicts of Interest 44
B. Cost Sharing or Matching 45
C. Other Eligibility Requirements 45
Section IV: APPLICATION AND SUBMISSION INFORMATION 45
A. Address to Request Application Package 45
B. Content and Form of Application Submission 46
1. Proposal Information 46
2. Proposal Preparation and Format 46
C. Submission Dates and Times 57
D. Intergovernmental Review 57
E. Funding Restrictions 57
F. Other Submission Requirements 57
Section V: APPLICATION REVIEW INFORMATION 57
A. Evaluation Criteria 57
B. Review and Selection Process 59
Section VI: AWARD ADMINISTRATION INFORMATION 60
A. Award Notices 60
B. Administrative and National Policy Requirements 60
1. Security Classification 61
2. Intellectual Property 62
3. Meeting and Travel Requirements 64
4. Human Use 65
5. Animal Use 66
6. Publication Approval 66
7. Export Control 67
8. Subcontracting 68
9. Central Contractor Registration (CCR) 68
10. On-line Representations and Certifications (ORCA) 68
11. Wide Area Work Flow (WAWF) 68
C. Reporting 68
Section VII: AGENCY CONTACTS 69
Section VIII: OTHER INFORMATION 69
A. Industry Day 69
B. References 69
1. Terminology Appendix 69
2. Example Tasks Appendix 77

Table of Figures:

Figure 1- Notional Concept of a Universal Reading System **13**
Figure 2 - Data Flows between Evaluation Team and Reading Teams **17**
Figure 3 - Hierarchy of Knowledge Capture **20**
Figure 4 - Reading Complexity Targets for each Phase **25**
Figure 5 - Text Readability Metric Description **33**
Figure 6- Performance Ratio Metrics Description **35**
Figure 7 - Effort Reduction Metric **38**
Figure 8 - Go/No-Go Thresholds by Phase **39**

Part One: Overview Information

- **Federal Agency Name** – Defense Advanced Research Projects Agency (DARPA), Information Processing Techniques Office (IPTO)
- **Funding Opportunity Title –** Machine Reading (MR)
- **Announcement Type –** Initial Broad Agency Announcement (BAA)
- **Funding Opportunity Number –** DARPA-BAA-09-03
- **Catalog of Federal Domestic Assistance Numbers (CFDA)** – 12.910 Research and Technology Development
- **Key Dates**
    - Posting Date – see www.fbo.gov
    - Proposal Due Date
        - Initial Closing – 12:00 noon (ET), 30 December 2008
        - Final Closing – 12:00 noon (ET), 13 November 2009
    - An Industry Day will be held on 21 November 2008 in Arlington, Virginia. See section VIII.A. below for details.
        - Interested parties must be registered by 20 November 2008.

- **Anticipated awards –** Multiple awards are anticipated for the Reading Teams and a single award for the Evaluation Team.
- **Types of instruments that may be awarded –** Procurement contract, grant or cooperative agreement.
- **Technical Point of Contact –**
    - Dr. Daniel Oblinger, Program Manager, DARPA/IPTO
    - EMAIL: DARPA-BAA-09-03@darpa.mil
    - FAX: (703) 741-0228
    - Defense Advanced Research Projects Agency/IPTO
      ATTN: DARPA-BAA-09-03
      Arlington, VA 22203-1714

Part Two: Full Text of Announcement

I. FUNDING OPPORTUNITY DESCRIPTION

The Defense Advanced Research Projects Agency (DARPA) often selects its research efforts through the Broad Agency Announcement (BAA) process. The BAA will appear first on the FedBizOpps website, http://www.fedbizopps.gov/. The following information is for those wishing to respond to the BAA.

DARPA is soliciting innovative research proposals in the area of machine reading (MR) in which candidate performer teams propose to build a universal text engine that captures knowledge from naturally occurring text and transforms it into the formal representations used by artificial intelligence (AI) reasoning systems. Proposed research should investigate innovative approaches that enable revolutionary advances in science, devices, or systems. Specifically excluded is research that primarily results in evolutionary improvements to the existing state of practice.

I.A. PROGRAM OBJECTIVES AND OVERVIEW

I.A.1. Background

Artificial intelligence (AI) systems continue to grow in use within the Department of Defense as there is a consistent emphasis on using high technology as a strategic advantage and reducing reliance on humans, both to save money and save lives. Situation awareness, diagnostics, prognostics, planning, logistics – all are areas in which AI systems are used or applicable. A great deal of the militarily relevant knowledge that these systems need is presently expressed as natural-language text. This knowledge may range from the local political and militant groups to infrastructure and food supplies. The necessary information is available, but rarely in a form that can be used by current AI systems.

For example, the U.S. military frequently faces impediments to stability and reconstruction operations in a new location due to the lack of understanding of the local situation. Similarly, strategic assessment of a foreign nation's science and technology base involves the continuous assessment of technical articles, bibliographies, conference agendas, etc. This information is often available on the World Wide Web, and some tools to assist this analysis are available, but the process would be significantly enhanced by a system that could directly analyze the information found in these text sources. The same reasoning could be equally valuable if applied to other types of open-source intelligence analysis, including assessing military readiness and posturing; political speeches, actions, and more obscure "messages"; economic trends and sentiments; and propaganda from terrorist groups and even their hidden web-based communications.

Looking beyond intelligence applications, there is a wealth of information contained in operations plans, charts from a commander's regular update briefings, after-action reports at all military echelons, lessons learned analyses, and similar textual documents

that are exploitable primarily by humans today. Diagnostic and trouble reports from fielded equipment contain a wealth of information that could be used for failure analysis and predictive maintenance, but only the simplest of analyses – often based solely on document metadata or term extraction – are automated today.

DARPA's envisioned Machine Reading Program (MRP) seeks to address the shortfalls discussed above. Currently, nearly all successful AI systems succeed because they possess sufficient consistent, relevant knowledge about a problem domain. However, since large amounts of knowledge are almost always needed for this success, AI systems require this knowledge to be expressed in a logical formalism of some type. Manually encoding such knowledge can become prohibitively expensive. Since text is, by far, the most flexible and ubiquitous medium used to capture knowledge about the diverse areas of human interest, it is natural to consider making it feasible for AI reasoning systems to employ this vast store of human knowledge. As AI systems currently cannot use such knowledge, it would be revolutionary if technology could be developed to bridge this gap.

The intent of the MRP described in this BAA is to enable just such a revolution. That is, the goal of the MRP is to create an automated Reading System that serves as a bridge between knowledge contained in natural texts and the formal reasoning systems that need such knowledge. However, rather than choosing a particular mechanism for this bridge, such as formalizing knowledge contained in natural text into a specific representation, the MRP focuses instead on applying the knowledge productively. As a consequence of this performance-based approach to the natural-language understanding problem, the MRP will only assess reading capability by its impact on some performance tasks. A relevant task could be a human performance task where a task-specific interface is employed to provide the user access to automatically read text, or it could be performance by a domain-specific AI system that requires information inferred by the Reading System from source texts. In either case, the goal is not the capture of (some or all of) the knowledge contained in a natural-language corpus in a general-purpose logical format; rather, the goal is the successful execution of a performance task in which knowledge contained in natural language is one essential component.

There is a long history of work in natural language understanding, knowledge-based systems, knowledge acquisition, and machine learning. All are areas that are potentially relevant to the MRP.

The Breadth of the Problem

There are enormous challenges in dealing with natural-language that have been well studied over the last 50+ years of research in automated text processing and natural language processing (NLP). Many are inherent in all natural languages (e.g., ambiguities at various levels), some are specific to individual languages (e.g., vocabulary, verb forms), some specific to the precise form of the source text, and some dependent on the (limited to comprehensive) level of understanding required to achieve the objectives of the system.

First, even within a single human language such as English, there are multiple modes of expression employed. That is, there is a tremendous variety of source texts that are, in some sense, naturally occurring. Text can be the medium for dialogs that are one-to-many, one-to-one, or many-to-many. Most documents meant for public consumption have well-defined styles, sometimes including a published style guide and human editors or copy editors. These include books (fiction, nonfiction, technical), technical manuals, technical or other scholarly journals and their papers, magazines, and newspapers. In addition, corporations, governments, militaries, and other organizations have private documents such as memoranda and email messages. The World Wide Web has led to newer formats that exist within a particular technical paradigm, such as web pages, blogs, and chat sessions. Other technologies, such as cell phones, have led to even more arcane documents and languages, such as the hyper-abbreviated lingo of instant messaging.

Second, naturally occurring speech can be converted to text and has many analogs to written document forms. Formal speech can include lectures, newscasts, and oral "articles". Less formal speech is found in phone conversations, in-person conversations, multi-person teleconferences, voicemail messages, etc.

An NLP system is typically built with one specific purpose in mind. Common purposes include information search and retrieval, content extraction, question answering, automatic translation from one natural language to another, and interactive natural-language dialog to support human-computer communications. To gain traction, previous systems have chosen either a breadth or depth strategy. With a breadth strategy, a system operates on a wide range of linguistic inputs, at the expense of the level of understanding achieved (e.g., search engines). The depth strategy provides a stovepipe of deep understanding, but only in a narrow area (e.g., with constrained vocabulary, subject domain, linguistic styles, and document styles). By contrast, the goal of the MRP is to develop broadly applicable technologies that can, with minimal customization or assistance, read to develop depth on their own in a variety of areas. To succeed at MRP performance tasks, a Reading System must be able to reuse a general reading capability to develop deep understanding of the content of a broad number of subjects expressed with a variety of writing styles.

Third, in addition to all of the many traditional NLP issues (both solved and outstanding), there are mismatches between the character and assumptions intrinsic to natural languages on the one hand, and to formal AI languages on the other. Bridging this gap will not only require dealing with the inherent complexities of natural language, but will also require the development of techniques for dealing with the inherent mismatches between the source text and the target's formal reasoning task.

One difference between natural languages and formal AI languages is that natural text does not employ a consistent, unambiguous way of naming objects, actions, events, relations, etc. By contrast, formal inference typically relies heavily on unique naming in order to inferentially connect facts in some performance task.

A successful Reading System must possess ways of compensating for this naming mismatch. It must identify mismatches between the natural and formal knowledge and bridge this gap to enable deep, task-specific reasoning over knowledge derived from naturally occurring text. Some possible approaches include:

- using adaptive matching heuristics,
- using feedback from task performance to inform matching,
- using convergent, statistical evidential reasoning about matching,
- maintaining islands of internally consistent mappings that are heuristically applied on a problem by problem basis,
- on-the-fly partitioning of language referents as dictated by a specific inference chain, and
- managing inferential closures as independent patches in order to minimize the impact of mismatched references.

This list is not intended to express a preference among approaches, nor is it in any way exhaustive. Rather, it is intended to indicate the scope of the problem that must be tackled.

Another mismatch is the implied universality of formal knowledge, versus the implicit domain-specific contextualization of naturally occurring text. For example, if one expresses the notion “everyone is here” straightforwardly in logic, the logical statement would imply that every person on Earth is in the same location. However, the English sentence “Everyone is here.” in a document is understood to include only those people who are appropriate, based either on an explicit definition of the set in an earlier sentence or an implicit definition that is provided more or less subtly, possibly requiring an understanding of many facts or hints from multiple places in earlier text.

Multiple incompatible representational ontologies can be used simultaneously in natural text, whereas reasoning systems assume a single, well-formed ontology applies throughout. To understand natural text, some form of meta-level reasoning must be employed in order to correctly apply the appropriate knowledge from the appropriate ontology. As an example of this very broad class of mismatches, consider one ontology of painting that specifies that painting is done by a person and a second in which painting is done by an air brush – without any characterization of the person doing the brushing. A Reading System must be able to determine which conceptualization is most appropriate to any given piece of text.

Reading is inherently ambiguous at many linguistic and logical levels, while formal reasoning is not. Some important types of natural-language ambiguity are word sense, anaphora (e.g., pronoun reference), part of speech, parse structure, and conceptual mapping. Consider the following examples:

- An example of word-sense ambiguity is the sentence, “You must adjust your attitude.” According to WordNet, “attitude” can mean a mental state or the relative orientation of a craft. Understanding the broader context in which this sentence occurs is the only way to determine which sense is correct.

- Consider: “Dave collided with Jim. He was unharmed. Jim, on the other hand, suffered a concussion.” The pronoun “he” refers to Dave, but one can only conclude this after reading the final sentence.

- Still more challenging would be the sentences “Mary replaced the silverware without thought.” versus “Mary replaced the silverware without designs.” This is an example of parse-structure ambiguity. The two sentences appear very similar in structure, but “without thought” is a mental state, which could only belong to Mary. Since it is unambiguous once the type of noun is considered, this is a well-formed sentence. On the other hand, “without designs” is very likely a descriptive term that applies to the silverware. Although it could conceivably also refer to Mary, word placement would strongly argue for the former interpretation. A good writer would move the phrase elsewhere if it were meant to refer to Mary.

- Consider a corpus of news releases that includes these two sentences: “In the Saarland the governing CDU was able to remain in power, the SPD lost seven seats, while the Liberals and Greens were able to re-enter state parliament.” “At the federal level, the Christian Democrats lost their dominance but remain part of the governing coalition.” Assuming that “CDU” is not in the reader’s lexicon, there are a number of ways to determine its meaning. One is to search the corpus (or other corpora) for more occurrences of “CDU” to see if it is explicitly defined somewhere. Barring that, one might be able to infer that “CDU” is an abbreviation for the Christian Democrat party by a deep conceptual understanding of each sentence, followed by a comparison.

- Human reading depends on making millions of “immediate” and implicit inferences that are so automatic that they are unnoticed when made. Consider an example involving the computation of frame axioms, immediate inferences about what assertions remain true after new information is read. When presented with the sentences, “Phil is broke. Phil is friends with Dominick. Phil is in his office. Phil is in Arlington.,” the reader would assume that Phil is still in his office (since being in Arlington is not inconsistent with being in an office). Subsequently reading the sentences, “Phil is now rich. Phil is friends with Ray. Phil is in the conference room creating a document.,” creates new default expectations that Phil is no longer broke; Phil is still friends with Dominick; and Phil is no longer in his office. Because these immediate inferences are implicitly assumed in natural texts and required to be explicit in formal reasoning, a full bridging solution will have some practical approach for addressing this disparity.

The classes of incompatibilities described above are clearly central to using machine reading (MR) of natural texts to drive formal reasoning tasks, but they are by no means exhaustive.

I.A.2. Expected Benefits

The overarching benefit of a successful Machine Reading Program will be enabling all knowledge encoded as natural text to be combined with the power of AI reasoning

methodologies, which will unleash a wide variety of new AI applications ranging from intelligent bots to personal tutors. For example, all of the text in the World Wide Web will become available for automating the monitoring and analysis of technological and political activities of nations; plans, rhetoric, and activities of transnational organizations; and scientific discovery within various disciplines. As digitized text from library books world wide becomes available, new avenues of cultural awareness and historical research will be enabled. With truly general techniques for effectively handling the incompatibilities between natural language and the language of formal inference, a system could, in principal, be constructed that maps between natural and formal languages in any subject domain.

The exploitation of convergent (mutually constraining) knowledge contained within one document or corpus promises to enable the development of NLP systems that perform significantly better than current systems and that read text much more like humans do. Perhaps counterintuitively, the use of more knowledge could potentially make a single disambiguation easier, by providing a wealth of constraining information to the analysis. Of course, the availability of a domain-specific reasoning system (DSRS) should add greatly to the constraints as well.

The MRP will also have made NLP deployable in a wide range of practical contexts, including almost arbitrary reasoning tasks fed by arbitrary types of text. Imagine if the same Reading System could be applied to scouring the World Wide Web for good deals on cars one time, and then applied to integrating new findings in genetics to an automated theory of disease. Another application might read newspaper articles and create a textual summary of the day's top news, prioritized by a person's own interests. All would use the same NLP front-end. Since MR may be viewed as a central AI problem, the entire AI field will also benefit – both during the MRP and after. Great progress is expected in NLP, domain-independent reasoning, and machine learning.

A successful MRP will leave behind a legacy of new researchers, new reading techniques and prototypes based on them, corpora, a variety of DSRSs, and evaluation methodologies; in other words, a focused, well-equipped, and enthusiastic research community. Many researchers beyond the initial MRP performers will build on the initial work of the program. An annual contest to evaluate Reading Systems from any source could be held to encourage further research and evaluate progress. In service of such a community, standard corpora could be made widely available. A central goal of the MRP is to foster a research community in machine reading. Consequently, as stated in paragraphs 1.17 and 1.18 of the Technical Volume description below, the Intellectual Property and licensing model will be part of the evaluation for this BAA.

I.A.3. Programmatic Approach

The programmatic approach taken by the MRP is to provide a research context that supports the widest possible variety of attacks on the general problem of building a bridge from natural text to formal AI reasoning. In performance terms, the MRP attacks the end-to-end task of employing textual knowledge in service of a variety of performance tasks. The MRP, as planned, covers five phases of 12 months each.

Each year/phase addresses increasingly complex and challenging tasks. Over the course of the program, the generality of the performance tasks supported will be progressively expanded. By working on easier aspects of the reading problem early in the program, a number of opportunities for transitioning the research to practical military and nonmilitary applications may happen sooner. For example, a highly proficient Reading System, informed by a DSRS that extracts explicitly stated facts from large volumes of text, could be extremely useful in any number of areas. A Reading System that extracts inferred facts would be even more valuable. A Reading System that could perform ontology extension or creation could even keep up by itself with the literature in a fast changing field such as genetics or international relations.

The desire for generality, and for separation of the reading and reasoning mechanisms from the task knowledge required, is built into the program structure. Therefore, there will be two types of performer teams: Evaluation and Reading. A single Evaluation Team will provide all of the domain-specific context required for reading and reasoning and work with the Government to conduct the formal evaluations. One or more Reading Teams will each provide a general reading mechanism that operates in these task-specific contexts. This program structure requires the development of a formal specification of a reading and reasoning task context. The Evaluation Team will evolve the specification outlined in this BAA to provide a plug-and-play software context that the Reading Teams will use as the platform for their Reading Systems. This context will be kept as approach-neutral as possible.

The MRP is structured to encourage approaches that employ general-purpose, domain-independent reading capabilities to the greatest degree possible. The program, through the use of growing numbers of challenge domains, seeks to frame this problem so as to preclude single-domain, fully custom solutions that do not translate to other application areas. Practically, performance evaluations will involve question answering, but queries are assumed to be coming from some DSRS. Since performance evaluations will be based on domain-specific reasoning tasks, general reading capabilities will still need to employ techniques like automated adaptation, heuristics, and machine learning to create islands of convergent (inferentially consistent) knowledge that are employed to achieve high performance on domain-specific tasks.

A variety of technical approaches are sought and expected in the MRP to address the MR challenges. A variety of approaches from within the field of NLP could be combined with approaches and techniques from many other fields. These could include deep knowledge approaches; unsupervised, semi-supervised, and supervised inductive techniques; and techniques that leverage domain-specific “tweaking” (of any kind) performed on a domain-by-domain basis by trained knowledge engineers or derived in some automated way from large existing knowledge sources. This list is by no means exhaustive, nor is it intended to express a preference by DARPA for one approach over another. Rather, it is DARPA’s intent to express the MR problem in a way that admits as many alternate approaches as possible. DARPA has no bias for or against any specific technical approach to either of the two areas (Reading or Evaluation) that may

be proposed. Each proposal will be evaluated on the strength of the case made for the ultimate success of the approach in the MRP.

The MRP casts the machine reading problem as bridging natural text and formal reasoning. This is a constrained formulation of the MR challenge but nonetheless embraces some very daunting challenges. Given this program context, success will not be measured by performance on some unconstrained Turing test based on the texts. Rather, the ultimate measure of success is the ability of Reading Systems to transform textual knowledge in such a way that it enables and improves the performance of multiple DSRSs.

In addition to the research and development directly funded by the MRP, DARPA also intends for this program to support and nurture a broader MR research community. DARPA plans to work with other government agencies to support system evaluation in an open competition that runs in parallel to or follows the MRP. Datasets and test cases developed for the MRP will be made available to the larger research community. DARPA expects MRP researchers to publish about developments in components of their Reading Systems as well as about system-level problem decompositions. Furthermore, this solicitation encourages proposals that will make components of the Reading Systems widely available. Whenever possible, these components, and the APIs developed for testing and for communication between components of Reading Systems (see section I.C.1.b), will be made available to all MR researchers regardless of their affiliation with the MRP.

I.A.4. Technological Goals

As stated, the goal of the MRP is technology that provides a bridge between the knowledge contained in natural text and formal reasoning systems. This section discusses the principal technical objectives that follow from that overarching objective.

- **A Universal Reading System:** Create a universal Reading System that can take any natural text and any reasoning context as input and can effectively apply the knowledge contained in that text in that reasoning context. Figure 1 below depicts such a system being applied to the problem of analyzing terrorists and terrorism events from open-source textual reports such as might be contained in Associated Press (AP) articles from around the world over a period of months to years. The Reading System is provided with pre-existing knowledge about terrorism accessed via a reasoning system focused on terrorism, a top-level query ontology (or Syntax Specification) that allows it to make queries of this knowledge and, finally, a large corpus of AP articles. By reading the articles to extract new information relevant to terrorism, the Reading System accumulates additional knowledge. The result should be enough knowledge to allow a high-performance reasoning system in a terrorism domain to perform successfully. Note that this high-performance reasoning system could either be a human (e.g., an analyst) or an AI system.

Also note that the Reading System will only be deemed universal if it performs adequately on a large variety of subject domains. A successful Reading Team will build only the universal Reading System. The pre-existing or background knowledge, query ontology and sample sentences that exemplify the ontology's concepts (Syntax Specification and Mapping), training and testing text corpora, DSRS, and evaluation queries with consensus responses will all be developed by the separate Evaluation Team. The Reading Systems will communicate with these external components through well-defined interfaces, which will be detailed in later sections of this BAA.

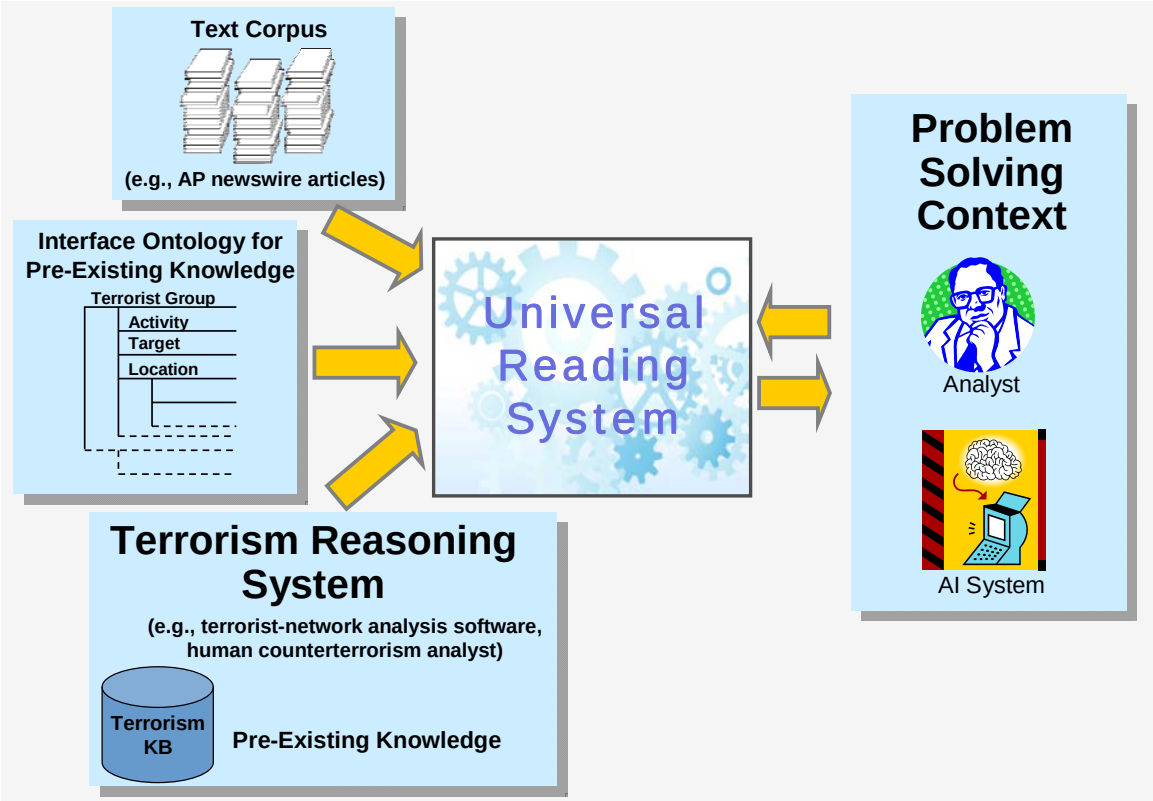

Figure 1- Notional Concept of a Universal Reading System

Figure 1 is not a functional specification; it is merely a notional diagram showing the principal inputs and outputs of a Reading System. For example, the overall system might operate in a data-driven fashion (i.e., it operates whenever it receives new text or a new problem domain), a purely goal-driven fashion (i.e., the DSRS asks the Reading System questions, which in turn causes it to look for relevant texts to process), or some combination. In addition, the various boxes shown may not actually be separate subsystems. For example, when the DSRS is an AI system, it will also provide the domain's pre-existing knowledge, and do so through the same interface.

- **Enhanced Capabilities from Combining NLP and AI Reasoning:** Combine NLP and AI reasoning into new technology that provides the benefits of both.

Today, reasoning allows a relatively small knowledge base to provide inferential power over a potentially infinite number of specific problem situations. Reasoning can use knowledge of many types and from many sources to create a robust solution for a specific problem. NLP can process natural text, the most ubiquitous source of encoded human knowledge. Making this knowledge available to reasoners should significantly increase the applicability of AI systems and reduce the cost of constructing them. The reverse is also true; NLP can benefit from formal inference over encoded knowledge. Thus, combining deep formal reasoning and NLP promises to strengthen both.

- **Handling Incompatibilities between Natural Language and Formal Inference:** As discussed earlier, explore general-purpose mechanisms to handle the specific classes of mismatches that occur between informal knowledge expressed in text and formal knowledge used for machine inference. Teams should propose techniques for managing these mismatches.

- **Domain-Specific Performance from a General-Purpose System:** Develop a general-purpose text-reading "front-end" (Reading System, for short) that can be used with any number of DSRS "back-ends." The program will explore mechanisms for semi-automatically obtaining high domain-specific performance from a general-purpose system. That is, it will investigate ways to create a high-performance Reading System constrained to use domain-specific text and a DSRS. Possible approaches may include (and extend beyond) the following:
    - adapting by learning new domains from small amounts of labeled data and large amounts of unlabeled text;
    - using large, general background knowledge bases that are not specific to target domains, but are used as part of an automated method for obtaining domain-specific performance across multiple domains; and
    - having a knowledge engineer provide domain-specific knowledge or tuning, in a time-constrained manner on a domain-by-domain basis.

- **Couple NLP and Convergent Knowledge:** The goal in coupling NLP with convergent knowledge – knowledge that is mutually constraining – is not to understand facts implied by a single sentence in isolation. Rather, the goal is to uncover physically remote facts, contexts, etc., and determine how these bits of knowledge can, in turn, help make the NLP feasible by constraining the interpretation of other such bits. Convergent knowledge could provide a source of power within the NLP context. Relating information drawn from different sentences, different corpora, or different sources (e.g., from text versus the DSRS) could provide valuable constraints for NLP itself. This program provides a context where such convergences are possible. Thus, exploration of how to best take advantages of this possibility is of interest.

- **Deployment of NLP in a Wide Range of Practical Reasoning Contexts:** Develop easily reused and re-purposed NLP technology that can be rapidly applied in a wide range of practical reasoning contexts where textual knowledge

could provide strong value. Enabling widespread application of this technology will require generality along two key dimensions:

- Text Format: handle broad classes of natural language in the formats found in situ (e.g., newswires, blogs, transcripts, web pages, etc.); and
- Reasoning Context: independence from the specifics of the reasoning context (i.e., can be developed without knowledge of the target reasoning tasks).

To the maximum extent possible, Reading Systems should also divorce the general reading mechanisms from the specific representations and reasoning formalisms employed in DSRSs. This somewhat unconventional approach is taken in recognition of the fact that the reasoning formalisms most auspicious for one reasoning problem and context are often cumbersome, or simply insufficient, for another reasoning task. Completely separating the Reading System from the DSRS may be impractical and unattainable. Nonetheless, the goals of generality and practical utility justify pushing this boundary as far as possible.

- **General-Purpose, Domain-Independent Mechanisms for Reading and Reasoning:** Explore general-purpose mechanisms for reading and associated reasoning that are divorced from the detailed background knowledge necessary for each domain-specific, high-performance task that may be assigned. As stated earlier, there is a common belief (borne out by research) that deep background knowledge will be the key for achieving high-performance Reading Systems. This program will explore the space of systems that employ task-specific background knowledge, representations, and reasoning paradigms, but in a way that separates them as much a possible from the reading mechanism itself. This separation serves the goal of rapid and widespread transition and adoption, as well as the important research goal of understanding what generalized capabilities contribute to reading performance across many reading and reasoning tasks.

- **Research Community in Machine Reading:** Another important goal is to nurture a research community (including and beyond those funded as part of MRP) that focuses on the MR problem. The MRP will pursue this in several ways:
    - Support on-going competition among multiple research teams, from the initial BAA responses and continuing with the evaluations at the end of each MRP phase. Intramural and extramural competition will spur broad interest in the topic and provide a vehicle for researchers to prove the value of their contributions.
    - Develop an evaluation methodology and quantitative metrics that provide ways to directly compare reading approaches.
    - Provide a framework for assessing the generality of reading approaches across multiple subject domains, by providing homogenized access to disparate reading task contexts, including text corpora and DSRSs.

- Where possible, provide ways to compare common components of a full Reading System in a way that allows a more fine-grained understanding of the tradeoffs involved in a large end-to-end Reading System. Support the specification of interfaces for a number of “plug and play” components, so that different instantiations of the same component can be swapped in and out for comparison testing.
- Provide “training data” and evaluation data for both the end-to-end reading and reasoning task and, where possible, the data needed to adapt and assess common subparts of the end-to-end task.

The MR problem is central to the future of AI. It is a large problem and many technical challenges will certainly remain unsolved at the end of the MRP. It is DARPA’s intent that many of the mechanisms established for the program will continue to be used by the research community long after the MRP itself has concluded.

I.B. Programmatic Details

Offerors should note that there is a Terminology Appendix at the end of this document in Section VIII.B.1 which provides definitions and/or greater detail on concepts and artifacts discussed elsewhere in this solicitation.

I.B.1. Team Types and Interfaces

As stated above, there will be two types of teams for the MRP: an *Evaluation Team* and one or more *Reading Teams*. As described in Section III.C.1 below, individuals and/or organizations that are performers on the Evaluation Team cannot also be performers on a Reading Team. See Section III.C.1 below for further details.

The Evaluation Team will produce multiple reading tasks and is responsible for the creation of the reading task domains, specification and development of the interfaces between the Reading and Evaluation Teams, and for the execution of the end-of-phase evaluations of the program.

Each Reading Team is responsible for producing an end-to-end Reading System that performs over all test tasks generated by the Evaluation Team. The reading tasks are described in greater detail later in this solicitation and examples are presented in the Examples Appendix (Section VIII.B.2.). Reading Systems will be built on technologies that apply in the context of tasks in which inference over textually-derived knowledge is required. Each MRP test set will be built on some implied performance task. Thus, the Reading Systems can be viewed as operating between text sources and AI (or human) systems in order to make the knowledge found in the text available in ways that enhance the performance of the overall systems.

Figure 2 below illustrates how the Reading Teams and Evaluation Team will interact programmatically by providing an overview of the data flows between the major components of the MRP. The bottom panel of the figure depicts artifacts produced by

the Evaluation Team: the *domain corpus*, the DSRS, and the *performance-task queries* used to assess the Reading Systems. The top panel shows the Reading Systems that are applied to each reading task specification and interfaces between the Reading Systems and the reading task artifacts that will be employed beginning in Phase 2. However, as a development aid, tasks expressed in this framework will be provided during Phase 1.

The Reading Systems will accept a natural text corpus and will process this corpus (parsing, mapping, indexing, etc.) as each team sees fit. A sequence of queries will be generated from some implied performance task (either a DSRS or human-in-the-loop performance). In either case, the Reading System answers the queries through a combination of reading the supplied text corpus and reasoning based, at least in part, on the task-specific inference and background knowledge available from the DSRS. Reading Systems will be evaluated on the correctness and completeness of their answers, as well as how much domain-specific knowledge engineering was performed by the Reading Teams for each task domain. Each Reading Team will be responsible for fielding a single end-to-end Reading System capable of reading in all task domains and responding to performance task queries.

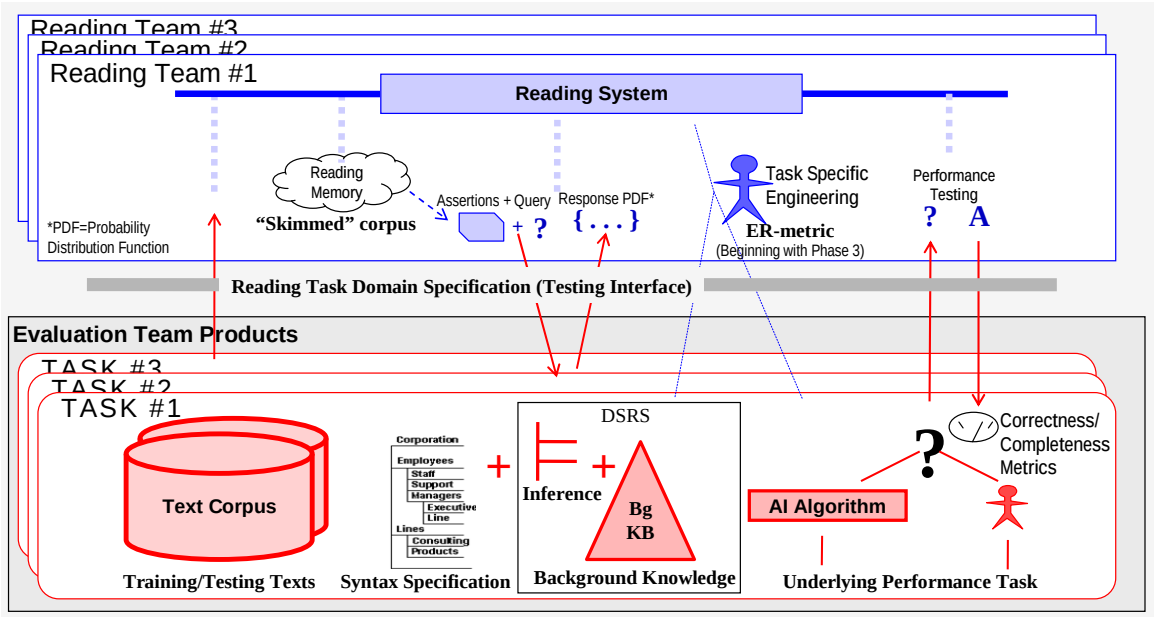

Figure 2 - Data Flows between Evaluation Team and Reading Teams

Each rounded rectangle on the bottom is called a *reading task domain*; these encode the complete testing context for a single application of the reading technology. Each reading task's components will conform to a common interface specification (the *reading task domain specification)* in order to simplify the testing and the construction of Reading Systems. Each task domain is constructed by the Evaluation Team as a combination of an input corpus, DSRS (including task-specific background knowledge), and a source of queries derived from an actual or simulated performance task.

Practical AI reasoning systems often adopt very incomplete, sometimes unsound, specialized reasoning mechanisms that are effective only on the target task, but are much less cumbersome and much more efficient than any general purpose logic mechanism (such as first order logic). Since the MRP goal is to develop general-purpose technology for mapping from a wide range of natural texts onto a wide range of formal reasoning systems, it would be inappropriate to select a single representation and reasoning system for the program. This program defines a crisp boundary, the *inference interface*, between the Reading System on one side and the DSRS – needed to understand the knowledge acquired by reading and to perform inference in support of performance tasks – on the other side. This interface hides many details from the Reading System, even details such as the *representational language* employed by the underlying inference. A Reading System may represent, index, and manage knowledge (including text, assertions, and inferential closures) in any way desired. However, this knowledge must ground out as (or be translatable to) statements in the target knowledge language.

This unusual separation exists partly to focus the Reading Teams on the most important part of the collection of reading tasks—the *text to knowledge translation*. The separation also facilitates assessing the general text-to-knowledge translation independent of the underlying task-specific problem solving. Practically speaking, rigidly enforcing this separation is probably impossible and would certainly be counter-productive. If the systems are to perform well for the more complex domains, Reading Teams will need to have some access to, and impact on, the DSRS. In these cases, the Evaluation and Reading Teams will operate more as collaborators in arriving at an effective ensemble system, while maintaining the generality expressed in the framework as much as each task domain allows.

The inference mechanism shown in the lower part of Figure 2 will be supplied by the Evaluation Team as a task-specific ***input*** to the Reading Teams, as part of the DSRS. At the beginning of the MRP, there may be no background knowledge or substantial inference capability provided; however, the mechanism will still be implemented to ensure that a Reading System can be used unchanged in later program phases that have simple or complex domains.

I.B.2. Program Structure and Phasing

As stated above, the MRP is planned as a five year effort in which increasingly complex and challenging tasks are addressed each year. Progression to each phase will be dependent on the teams passing the previous phase's Go/No-Go targets, funding availability and other program considerations. Each year, the Evaluation Team will be challenged to produce larger and more complex reading tasks and, each year, the Reading Teams will be challenged to address these tasks and to demonstrate increasing performance against the program's metrics on both new and previous-year reading tasks. Each year, the Go/No-Go performance targets will increase as will the number of problem domains used in evaluating the Reading Systems. At the same time, different dimensions of qualitative complexity will be emphasized in each phase.

At least one reading task in each program phase will push the boundary on the qualitative dimensions being exercised, although every domain will not push every boundary in reading complexity. Reading Systems will be expected to achieve Go/No-Go targets on new domains each year, as well as on domains from previous years that are retained in order to provide diversity and ensure the generality of the systems. This approach is intended to enable end-to-end testing of the Reading Systems starting in Phase 1, with controlled increases in difficulty in each subsequent phase.

The text in the corpora for the reading tasks will be English language text characteristic of naturally occurring text. That is, text found “in the wild” will not be rewritten for the program in order to meet artificial linguistic constraints. In each phase of the program, the Reading Systems will be required to extract knowledge that is expressed with more complex language constructs (e.g., time and modalities) and involving more complex concepts than in the previous phase. However, when constructing a new problem domain, particularly during the early phases of the program, the Evaluation Team will work to ensure that the text corpora contain ample content written at a level appropriate to the program phase in which the domain is introduced.

A programmatic goal of the MRP is to produce and evaluate an end-to-end Reading System in each phase of the program that progressively takes on increasing fractions of the full complexity of the problem of bridging natural and formal knowledge forms. As described in the Metrics section (I.C.1.c.) below, reading systems are expected to achieve higher performance targets at each phase and to do so with progressively more challenging reading tasks. Thus, in addition to the increasing quantitative performance goals, it is important to circumscribe qualitative aspects of the bridging task as well in order to make the overall goal plausible. Natural text will often contain high degrees of complexity in all phases of the program (this is the nature of truly natural text). The classes of text that the teams are responsible for productively processing will be scoped larger in each phase.

Although developing a fully general Reading System is a daunting challenge, it is important that the MRP measure progress toward the overall objective. Many large research programs have built separate components in a bottom-up fashion for a number of years, only to find that they don’t solve the overall problem when integrated. Therefore, the formal evaluation at the end of each program phase will test the performance of a general-purpose system, rather than testing its subsystems separately. To this end, the MRP is designed so that each phase carefully generalizes the capabilities and relaxes the restrictions of the previous phase, along various dimensions.

In this view, the MRP aims at the easiest and highest-payoff aspects of knowledge capture from text first, only then moving on to the more challenging aspects. Figure 3 below shows one aspect of this progression as a pyramid with easier to acquire knowledge toward the bottom.

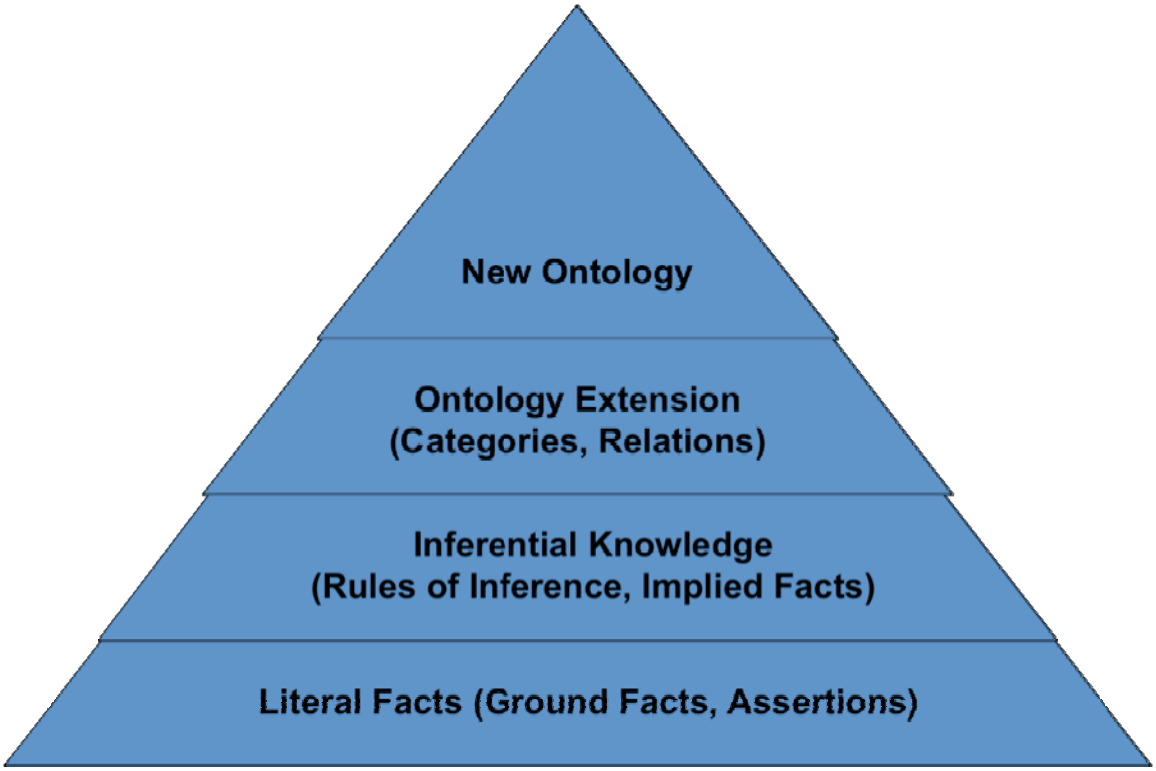

Figure 3 - Hierarchy of Knowledge Capture

At the base of the pyramid is the capture of literal facts from text. This assumes there is already, in the DSRS, a knowledge representation structure, known background knowledge, and available reasoning mechanisms. There is high value in capturing asserted facts because many practical applications require a large number of such facts. The capture of this knowledge is also somewhat easier because the existing knowledge of types higher in the pyramid heavily constrain this knowledge.

The next layer of the pyramid, at least in volume if not in difficulty, is the capture of inferential knowledge. This layer is also constrained by the representation provided by that reasoning context. Inferential knowledge refers to knowledge that controls the derivation of literal facts from other literal facts, both at the symbol and knowledge level. That is, it controls what follows in a logical sense and which inferences are and are not auspicious to make in any given context.

Topping the pyramid are the two related challenges of extending and creating the ontologies for a domain. The lower of these addresses extending an ontology (e.g., by adding new subcategories, new relations, etc.). The upper layer is the least constrained challenge and represents the capture and creation of knowledge from text, namely, that of a totally new ontology.

Four dimensions of complexity associated with reading are addressed below, along with our current understanding of each of these dimensions. The dimensions of complexity are:

- *Translational* – the complexity associated with translating from text to individual assertions;
- *Inferential* – complexity associated with combining groups of assertions in order to obtain other assertions;
- *Representational* – complexity associated with the syntactic and semantic complexity of the target representation of assertions themselves, and complexity of the corresponding linguistic forms associated that are mapped onto these assertions; and
- *Scale* – complexity associated with the total number of assertions being managed (both in the natural and formal representational forms).

Translational Complexity

This is a qualitative measure of the difficulty of translating text into individual formal assertions, or "chunks" of knowledge. The lowest levels of translational complexity involve a nearly one-to-one mapping between relevant parts in the textual and formal representations and little variation in the language used to express the assertions in text. High levels of translation complexity involve significant variability in the natural text, missing or extra information, and ultimately unspecified intermediate representational constructs. More specifically, this translational dimension is expected to grow from direct to generative translation along the lines below:

> **Direct –** A direct translation is one in which a single sentence in the source text contains the fields of the target formal assertions that are to be captured. This allows for missing fields (for example an assertion that something was purchased might include a price field that could be left blank if the sentence did not specify it.) If additional fields can be gleaned from other sentences, they are ignored when assessing direct translation. Further, we assume that all target syntactic forms, relations, and fields are specified prior to reading.
>
> **Variable –** As with direct translation, a variable translation is one in which a single sentence in the source text contains all the fields of the target formal assertions to be captured. The difference is that the surface forms encountered are quite different than those forms provided in the Syntax Specification to sentence mapping examples.
>
> **Extensional** – Adds the requirement that the system is capable of extending class and relational hierarchies, as well as adding new fields to each as needed by the target reasoning task.
>
> **Generative** – Adds requirement that entirely new classes and relations must be derived from the supplied texts in order to address the target performance tasks. These classes and relations will be drawn from the text, and depending on the level of inferential complexity will participate in inferences that stretch from the source text to those syntactic forms supplied prior to reading.

Inferential Complexity

Inferential complexity is a qualitative measure of the control necessary for guiding the chain of inferences that determine the derivation of new assertions from existing ones. This involves the acquisition of inference "rules" from text as well as the determination of which inferences should be made. In early program phases there is no inferential complexity required beyond that supplied by the entailment operator. Later phases will require induction of increasingly complex rules for chaining assertions that will be required within each performance task. Thus, inferential complexity is a measure of the inferential knowledge that must be acquired and applied in order to address the reading task. Teams will propose the forms of inferential knowledge they will acquire from text and the approach(es) they will investigate. This inferential dimension is expected to scale along the lines below, from an initial position in which everything is provided by external sources to a level of complexity that requires inferencing over novel classes and relations that are discovered from reading.

Provided – Early phases of the program do not require acquisition of inferential knowledge. Relevant inferential knowledge is supplied prior to reading. Further the supplied inferential knowledge is "self limiting" in the following sense: given any set of assertions, K, and null query template (which matches all consequent assertions) the inference mechanism is guaranteed to output a set, K', assertions that contains K, and the size of K', |K'|, is some small multiple |K|.

Inferential Control Knowledge – This is knowledge (expressed differently by each team) that controls when and which inferences to perform, and which assertions drawn from the text corpora to employ. Practical AI systems have domain-specific solutions for controlling inference. Given that the MRP is bridging from natural to formal knowledge, it must also bridge this missing control knowledge if the target reasoning system (the DSRS) is to be effective for the target performance task. It should be noted that one need not "read" this control knowledge explicitly from the text (as it will often not be found in the text). Using statistical learning, specialized reasoning or other techniques the teams must make effective use of the textual knowledge in the specified targeted performance context. The Examples Appendix at the end of this document provides a range of such performance contexts for consideration.

"Immediate" Inferences – Immediate inferences are a loose category of inferential knowledge that surrounds each new relation and categorical term in some domain-specific ontology. These classes of inference rules must be acquired automatically because manually encoding them would be too costly, as families of these rules must be encoded for each new relational term and category considered. In later phases, teams will recommend the syntactic structure of this knowledge—they will need to balance its generality verses the complexity of its acquisition. Two likely example classes of such inferential knowledge might include frame axioms, and data consistency rules. So, for example, consider that if a person is in the kitchen then later, in the living room, we know they are no longer in the kitchen. However, if instead, we later find they are in New York, we cannot assume they are not in

the kitchen. Also, if my eyes are brown, then they cannot be blue, but if my friend is Bill, Jack could be my friend as well. Such immediate inferences are large in number, and are essential in understanding text. Thus, a general purpose Reading System must somehow acquire such knowledge.

Inference Over Novel Structure – The task of acquiring inferential knowledge from text is more challenging than acquiring assertions because it is less constrained by any given piece of text. In the final phase of the program, Reading Teams will be handling *generative* translational complexity as described in the previous paragraphs. This means that the target representations will include newly acquired relations and categories. Mixing that complexity with immediate inference acquisition will add complexity to that task. It is already under-constrained even when building from a predefined syntactic structure. Finally we will be testing the acquisition of immediate inferences over acquired structures; in this final case one must simultaneously deal with uncertainty over appropriate syntactic forms as well as appropriate generalizations of patterns read.

Representational Complexity

This is a qualitative measure of the representational power of the target knowledge representation (the syntactic and semantic complexity of the assertion language), as well as the complexity of the linguistic constructs that the system must correctly transform into this target knowledge representation. In essence, it refers to the intrinsic complexity of the knowledge chunks (assertions) that are captured from text and must be reasoned with. In the early phases of the program, systems will be responsible for capturing and reasoning over timeless and modeless asserted relations between entities (see discussion of assertion representation in the Terminology Appendix). In later phases, the systems must capture and reason over temporally modified English and temporally constrained assertions. Systems will also begin with a modeless representation of the world, but will then progress to deal with modal language as well as the back-end mode-ed assertions. For clarity, an initial Assertion Representation as Resource Description Framework (RDF) triples will be provided. This representation will be extended over the course of the program to capture the necessary representational complexity as the scope of the program expands. As was done for translational and inferential complexity, below we define the lowest level of this representational dimension followed by an indication of how it will be extended over the program.

Simple Relational Model – In the first phase of the program, we focus on the acquisition of simple categorical and relational knowledge. The Syntax Specification described in this section specifies a method for defining a space of categories and typed relations over those categories. The *simple relational model* defines assertions as a set of relations over instances from the specified categories. These assertions are taken to be timeless, to apply in all contexts, and to have no modal qualifications.

Temporal

Temporally Qualified Knowledge – The first extension of this basic model is to allow for assertions to apply to absolute time points and intervals.

Relative Temporal Qualification – This basic capability is extended to handle relative temporal qualifier, and faithful translation from natural text. For example, “wait until you have a first draft of all sections before writing the introduction”.

Modal/Contextualized Knowledge

The full complement of language mediated and semantic modes used in all of natural language extends beyond the scope of this program. Indeed we first strive for generic mechanisms for binning assertions from text into “contextual buckets” that support simple methods for controlling inference over those assertions. Throughout the program we increase our support for more faithfully handling the semantics of frequently used buckets, as well as support for the linguistic forms associated with those modes.

No Modal Representations – Even in this simplest configuration teams cannot ignore the existence of modal language. If the system reads, “John wishes he had a million dollars” it must know enough not to add <has John MillionDollars> as an assertion of fact about the state of the world.

Generic Modal Processing – The basic model is extended to include support for reasoning within contexts and basic support for translating “lifting” knowledge from one context to another. Here we blur predefined linguistically mediated contexts like “John wants …” and domain-specific contexts defined by text structure and proximity. For example, the following sentence in a section about two-stroke engines must correctly separate knowledge about prototypical four-stroke engines. “… unlike the four-stroke engine which completely clears the combustion chamber before injecting new fuel, the fuel is injected into a mixture of fresh air, and exhaust gasses.” Before we can correctly handle the semantics of these myriad linguistically and semantically defined contexts, we must simply be able to recognize the boundaries between the contexts, correctly bin assertions by context, and handle context- or mode-specific “lifting” of assertions between contexts.

Semantic Model Processing – For some fixed set of the most useful modal language and semantic contextual classes we will take on the linguistic translation, formal representation, and inference to processes these forms. This could include belief and intention, as well as possible consequential worlds and intermediate states during action execution, etc.

Scale Complexity

Scale complexity is perhaps the most straightforward of these dimensions. Phase 1 of the program will use hundreds of relatively homogeneous texts while Phase 5 is

expected to use millions of pages of text comprising multiple writing styles conveying relevant, irrelevant, and often conflicting information.

The MRP will not be driven specifically by the complexity contained in natural language. Instead, we will increase the complexity of the implied performance tasks underlying the queries given to the system. These increases in complexity will guide the complexities of natural language that are taken on and, perhaps more importantly, will guide the degree of faithfulness needed from each Reading System in mapping onto those representations.

Specific targets are listed for each of these dimensions in each phase of the program (see Figure 4). These targets were chosen to (1) distribute the burden of the full complexity of MR over the life of the program, (2) synchronize the advances on each of the dimensions in order to allow them to build on capabilities developed on other dimensions in a previous or current phase, and (3) develop those aspects of the full MR tasks that have earlier transition opportunities.

The precise meaning for each of the defined levels for these dimensions is expected to evolve as our understanding of the program evolves. Indeed, even the dimensions themselves may well evolve as our understanding of the MR task grows during the program. Once teams have been selected, the program will undertake to clarify these dimensions to include or exclude categories of knowledge and linguistic forms, as appropriate for the approaches under consideration. Stylistically, the dimensions are discussed below without further repetition of the disclaimer that the details are expected to evolve.

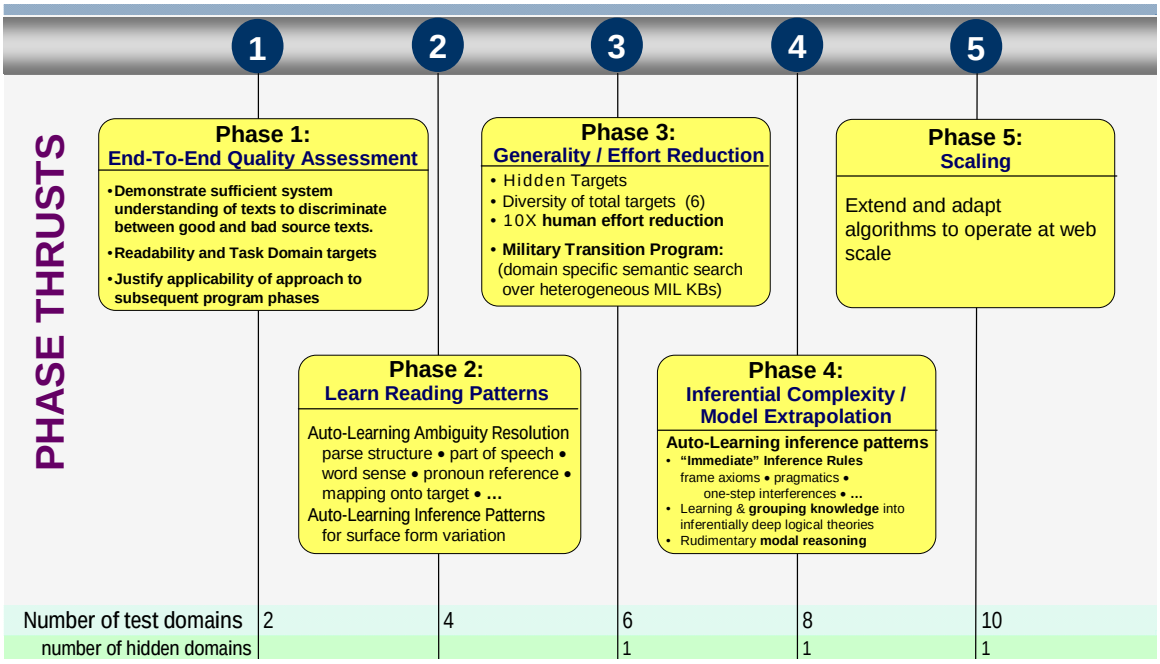

Figure 4 - Reading Complexity Targets for each Phase

The characterization of the dimensions of reading complexity makes as few assumptions as possible regarding the details of the reading process. However, the characterization and discussion does assume that the target representation is composed of knowledge chunks (called assertions) and the dimensions are defined around this unifying assumption. N.B.: It is understood that while many, though not all, targets for MR would satisfy this assumption, it is made here in order to quantify the reading task.

I.B.3. The Focus of Each Phase

The early phases of the program focus on the acquisition of facts (assertions) from text. In later phases, the focus shifts to the acquisition extension and creation of new formal representation and of acquisition of inferentially generative knowledge ("rules") from text. Teams should propose the classes of "rule" knowledge they will acquire from text and describe their approaches for building these capabilities. This rule extraction is expected to require balancing the generality of the acquisition engine against the practicality of its execution. The object is to automate the acquisition of those portions of the background knowledge that would be very costly to manually encode on a task-by-task basis. The creativity and credibility of approaches presented in this aspect of the proposal will be a factor in determining the merit of a proposal.

The rest of this section describes the qualitative dimensions of change that will drive the challenge levels higher each year. There is a fundamental qualitative distinction between the Phase 1 goal and those that follow. In particular, Phase 1 addresses the problem of assessing the readability of text, which is presumed to be a precursor for solving the more general readability problem. The rationale for this approach is described below.

Phase 1

Since the program is predicated on the idea of extracting inferentially consistent knowledge from the text provided, it is important that the system be able to assess what "good" text is. The quality of text found "in the wild" varies considerably with source and author. Thus, systems must be able to assess the quality of the text provided as input. During Phase 1, the primary focus will be on developing end-to-end systems that are able to understand enough of the text to provide an overall assessment of the quality of the text. The Go/No-Go performance metric employed for this phase assesses the ability of the Reading Systems to identify passages that are clear and unambiguous and thus accessible sources of knowledge.

Attempts to extract knowledge from poor quality text are likely to prove futile, and accepting knowledge derived from such text into its model of the world will result in a model that cannot be used effectively for inferential purposes (which is the ultimate goal of the program). Thus, the first phase of the program will focus on assessment of text quality itself. It is anticipated that deep assessment of text quality will require many aspects of the full Reading System. Therefore, Reading Teams are expected to design their Phase 1 system in a way that aligns it with the longer term goals of the program.

In Phase 1, the Reading Systems will accept a set of sample text documents from the Evaluation Team and must assess the relative quality of each text. Naturally occurring texts are not all of high quality, and incorporating even small amounts of incorrect knowledge into an inferential context has the potential to greatly degrade the operation of the Reading Systems. Reading Teams will be designing and building an end-to-end Reading System but, in Phase 1, the Go/No-Go evaluation will be based on each system's ability to assess text quality.

Reading Systems will need to solve many of the challenges of the overall program goal of transforming natural language text into inferentially useful forms in order to accurately classify the readability of documents. The Text Readability Metric will be used in Phase 1 to compare Reading System performance to human performance on the task of identifying clearly written, unambiguous text. This metric will assess the clarity of information content in text passages by comparing the rankings of passages by both humans and computers. All metrics and the measurement procedures are discussed in more detail in Section I.C.1.c below.

Reading Teams that successfully pass the Phase 1 Go/No-Go performance metric will have demonstrated a robust ability to screen texts of low quality. Additionally they are expected to have other components needed for an end-to-end Reading System. As subsequent program phases require end-to-end Reading Systems, Reading Teams must provide compelling evidence that they are on-track to meet the requirements of subsequent program phases. Additionally, because the metrics for later phases differ from the Phase 1 metric significantly, teams will need to demonstrate (or otherwise justify) their competence at taking on the later phase metrics *at the end of the first phase.*

The Evaluation Team will produce and provide test reading tasks which could be employed in this justification during Phase 1. Reading Teams may use reading task domains produced by the Evaluation Team, or other means, to justify that their technologies will successfully apply to the Phase 2 and beyond Go/No-Go performance metrics. Reading task domains produced by the Evaluation Team for this purpose will be smaller in size, and attempts will be made to control the variability in those texts, though no editing or simplifications of the text will be done—they will be naturally occurring texts.

Phase 2

Phase 2 of the program will focus on building end-to-end Reading Systems that handle the variability that exists in natural texts (the second step in the earlier discussion of Translational Complexity). For example, Reading Systems will be expected to extract knowledge from texts that requires handling complexities like intersentential pronoun resolution and anaphora, use of synonyms, contextual part of speech resolution, active and passive voice, variations in sentence structure, etc. Domain corpora will be larger, comprising roughly a thousand documents. Systems will be expected to demonstrate initial capabilities for answering queries that require inferential processes using the target domain description and extracted facts. The Go/No-Go performance metrics

employed for this phase will compare the effectiveness of the Reading Systems to that of humans in answering queries. The effectiveness of the Reading Systems at extracting facts and using them to answer fact-based queries will be assessed. Reading Systems will also be expected to demonstrate the ability to answer relatively simple queries that are not strictly fact-based, but instead require inferences valid within the domain ontology.

Phase 3

Phase 3 will focus on the generality of the Reading Systems and their efficiency in extracting knowledge from text. Domain corpora introduced during this phase will comprise a few thousand pages. Beginning in Phase 3, one new domain will be hidden from the Reading Teams until Go/No-Go testing begins. This use of hidden domains is intended to assure that Reading Systems are not tailored to specific problems but can, in fact, be readily adapted to any domain. Along the Representational Complexity dimension, more complex conceptual representations (e.g., temporal reasoning with definite and indefinite durations) will be required during this phase. The systems will be required to show improved performance over Phase 2 for fact- and inference-based queries in order to meet Go/No-Go expectations. The Effort Reduction metric will be introduced as a Go/No-Go measure during this phase. This metric compares the human effort required to adapt a Reading System to a new domain with the level of effort that would be required to manually populate a knowledge base with domain knowledge.

Phase 4

Phase 4 will require advancement along all the dimensions of complexity and will focus on significantly extending the inferential and modeling capabilities of the Reading Systems. Accurate knowledge extraction and representation during this phase will require systems to acquire and employ simple pragmatic inferential knowledge, properly acquire and apply immediate inference rules, accommodate rudimentary modal reasoning, and learn and group knowledge into inferentially deep logical theories in order manage read knowledge and control inference. Domain corpora introduced during this phase will be comprised of tens of thousands of pages and will contain inconsistent and contradictory statements. The systems will be required to show significant performance improvements over the Phase 3 goals for both fact- and inference-based queries in order to meet Go/No-Go expectations. Similarly, an order of magnitude increase in performance against the Effort Reduction metric will be required of the Reading Teams in this phase.

Phase 5

Phase 5 will aggressively push the computational efficiency of the algorithms that have been developed in prior phases of the MRP towards handling web-scale input corpora. Domain corpora will be “web-scale” (e.g., hundreds of thousands to millions of pages). Reading Systems will be expected to automatically extend the domain ontologies to include new concepts discovered from the source texts and queries will include concepts found in the texts rather than the original domain specification. The systems

will again be required to show significant performance improvements over the Phase 4 goals for both fact- and inference-based queries (particularly the latter) in order to meet Go/No-Go expectations. Another order of magnitude increase in performance against the Effort Reduction metric will be required of the Reading Teams in this phase.

I.C. Performer Expectations

Offerors submitting bids as Reading Teams should carefully review the Evaluation Team expectations (I.C.2) as well as their own since a number of the program artifacts that will be used by the Reading Teams are deliverables of the Evaluation Team and are described in that section. By the same token, offerors submitting bids for the Evaluation Team should carefully review the Reading Team expectations in order to fully understand the context of their own requirements.

I.C.1. Reading Teams

I.C.1.a. Task Details

Expectations Over All Phases

- Support the Evaluation Team in developing, specifying, and revising:
  - Query interfaces and response formats,
  - The inference interface, which provides access to the DSRS,
  - Common intermediate interfaces within Reading Systems.

Phase 1 Expectations

- Develop/Demonstrate a Reading System that accepts domain corpus documents as input and classifies them according to their clarity and level of ambiguity. Reading Systems shall perform at least as well as human performance measured by the Evaluation Team in identifying clearly written, unambiguous documents (see section I.C.1.c.).
- At the end of the phase, use a prototype or other demonstration to show that a complete end-to-end Reading System that is capable of meeting subsequent Go/No-Go criteria, will be delivered at the end of Phase 2.

Expectations for Phases 2-5

- Deliver a Reading System and demonstrate end-to-end performance with two new (in each phase) reading task domains provided by the Evaluation Team, as well as with reading tasks developed during previous program phases.
  - Achieve Performance Ratio for Facts (PR_F) goals shown in section I.C.1.c for all reading tasks, including reading tasks introduced during the previous phase (possibly updated as the program evolves).

- Achieve Performance Ratio for Inference (PR_I) goals shown in section I.C.1.c for all reading tasks, including reading tasks introduced during the previous phase (possibly updated as the program evolves).
- Beginning in Phase 3, achieve Effort Reduction goals shown in section I.C.1.c for the hidden reading tasks introduced during each phase.

I.C.1.b. Performance Expectations and Deliverables

In Phase 1 of the MRP, a Text Readability metric is used for the Go/No-Go assessment. In subsequent phases, measures of correctness/completeness (the two dimensions are integrated into a single f-measure indicator) and effort reduction are employed in Go/No-Go assessments. The specifics of the Metrics are described in greater detail below.

Achieving the Go/No-Go targets presented in Section I.C.1.c for each phase of the program is the basis for establishing a team's eligibility for continuing to work on subsequent phases of the MRP. Other factors include funding availability and other program considerations.

Over the course of the program, Reading System development may be guided by objectives such as interoperability and system generality. Such considerations are elaborated in this section. The MRP makes as few assumptions as possible about the structure of possible Reading Systems. Those assumptions that *are* made are oriented toward providing a common input and testing framework.

Some common intermediate data structures and reading sub-processes that cross team boundaries may emerge as the program progresses. In those cases, teams will be encouraged to adapt their interfaces and approaches to accept and employ these intermediate results. DARPA is pursuing this in order to:

- Go beyond assessing the end-to-end performance of systems against each other to also assess the relative performance of some sub-components against others, regardless of the source of those components. This is important for a scientific understanding of those components as well as for the understanding needed to drive the performance of those components within the program.
- Capture enhanced datasets (e.g., with intermediate data such as annotated parse trees) produced from running the end-to-end systems over training and testing corpora. These enhanced datasets with associated intermediate results, can be used within the MRP *and by others* to support research on specific sub-components of a longer MR pipeline. A growing set of rich correlated datasets, all in common formats, will greatly aid empirical research over multiple areas within the MR space.

In subsequent phases, Reading Teams could opt to focus on more advanced aspects of the MR problem if they establish that components from other Reading Teams are sufficient for less challenging aspects of the reading problem.

The MRP does not intend to dictate these intermediate interfaces, mandate the level of commonality to be achieved, or force all teams into compliance with such interfaces. Instead, the Government charges the *Evaluation Team* with the job of proactively working with the Reading Teams to identify commonalities and, as the possibility for agreement arises, to find ways to provide artifacts, datasets, and interfaces for all teams, and the larger MR research community, to exploit as they see fit.

In contrast to the requirements and expectations of the Evaluation Team, far less is specified in this BAA for the Reading Teams. This is intentional, as our goal is to provide as wide a platform as possible for alternate reading approaches to be considered. Figure 2 above illustrates that each Reading System is expected to accept reading tasks as described above. DARPA believes that some teams may wish to perform parsing, indexing, and deep processing of these texts immediately upon receipt of the texts and the target Syntax Specification. The artifacts of such preprocessing are denoted by the *"skimmed corpus"* placeholder in the figure. These intermediate forms are used in some way specific to a Reading Team in order to answer the supplied stream of problem queries.

The DSRS provided for each reading task will also be used by each Reading Team during this process. Accessing the DSRS via the inference interface provides a way to determine the set of assertions that are entailed by some other "given" set of assertions, though it does not specify which assertions should be employed for any given problem query. This is intentional. Simply parsing all texts into a single, large undifferentiated set of assertions, and then applying all available inferences, would be expected to produce an inferential closure that is hopelessly and uselessly inconsistent. Thus, a key challenge for building a robust bridge from natural languages onto formal languages will be determining *which* set of assertions – potentially drawn from a very large superset extracted from large source corpora – should be employed within the context of any given inference chain.

While Reading Teams may supply and employ their own "common sense" repositories, the task-specific background knowledge in the DSRS described above is intended to capture relatively narrow (and internally consistent) theories required to perform the reasoning required for the supplied particular reading task. In the early phases of the MRP there will be little or no background knowledge supplied, as the reading tasks will be more focused on the challenge of learning to map basic facts from texts to reasoning contexts. Later phases will employ the inference interface to construct reading tasks that are embedded in complex AI reasoning contexts, without requiring the Reading System to "read" all the necessary reasoning capabilities.

This solicitation imposes no constraints on how texts are parsed or on how subsequent knowledge and inferences are managed and indexed. This leaves room for diverse approaches to the basic problem of bridging between natural language and formal knowledge. The only constraint is that whatever structures are devised for managing text, assertions, and inferential closures, they must reduce to (or be translatable to) the

RDF representation of assertions and the probabilistic entailment expression that conveys which assertions inferentially follow from others.

The Reading Teams are encouraged, whenever practical, to generalize their systems away from the specifics of any given performance domain by developing algorithms and heuristics that exploit the Syntax Specification to capture task-specific knowledge representations rather than hand-tuning or training the system to a domain. The probabilistic entailment operator is intended to provide a mechanism for pulling task-specific inferential power back into the reading process itself. For example, a task-specific inference rule in the DSRS might assert that a person usually works for the same company as their manager. A general Reading System might hypothesize that the sentence "John recently left LargeCorp to work for DARPA" implies <WorksFor John DARPA>. Domain-specific inference could be employed in a general way to query whether convergent knowledge from other sources supports this assertion. For example one could query <WorksFor John ?x> given assertions derived from text: {<Manages Bob John>, <WorksFor Bob DARPA>}.

The goal is to decouple the Reading Systems from domain-specific inferences in a way that allows such knowledge to be used in adapting and applying the Reading System to specific task domains without keeping separate the Reading System from such domain-specific details. While it will not be possible to completely decouple the general and domain-specific portions of the Reading Systems, early performance benefits (in the Effort Reduction metric, for example) are expected. Further, we expect practical application of this technology to be enhanced by our attempts to distill the task independent aspects of the technology.

This may be a task that human operators are engaged in (e.g., searching for the cheapest single source for a basket of products), or it could be a task performed by some AI algorithm (such as matching a template against a stream of crime reports for emerging behavioral trends). In either case, this activity will result in a set of formal queries that must be answered based on the text. Accurately responding to these queries is expected to require building models of the text that connect different parts of the test corpus into a single inferential chain that addresses the performance task. This second requirement—that the inferences require consistency over large sections of text—distinguishes MRP tasks from those for which existing document retrieval or question answering approaches would apply.

Just as the probabilistic entailment operator is expressed in general terms, the queries posed by the performance task will be expressed in similarly general terms. Queries that are part of a performance task will be issued using the same query template employed by the inference interface and the result will be a set of probabilities over ground instances just as with the inference interface. The only distinction is that the entailment operator explicitly accepts a set of assertions that serves as the basis of inference. The performance task only supplies queries; the entire test corpus is the informational basis for answering the queries. A non-trivial challenge for the Reading Systems will be to decide *which* parts of the texts and which text-to-knowledge

translations should be used as the basis for each query. Systems are anticipated to employ measures of internal inferential consistency, contextual clues of relevancy, and many other techniques.

In addition to the required reports described elsewhere in this BAA, Reading Teams shall deliver systems capable of passing the Go/No-Go criteria at the end of each phase of the MRP. In Phases 2 through 5, this shall include delivering complete Reading Systems (to include source code, libraries, software modules, executables, etc. that are consistent with the proposed intellectual property and licensing models) to the government at the end of each phase.

I.C.1.c. Metrics

I.C.1.c.i Readability

The Text Readability Metric compares the text quality assessments of humans and computers. The average difference between human ratings and the consensus rating for a passage of text defines a performance threshold for that passage. The Phase 1 Go/No-Go criterion is that the machines' average agreement with the consensus ratings is better than that achieved by humans – in other words, on average the machines exceed the performance threshold set by human text rating.

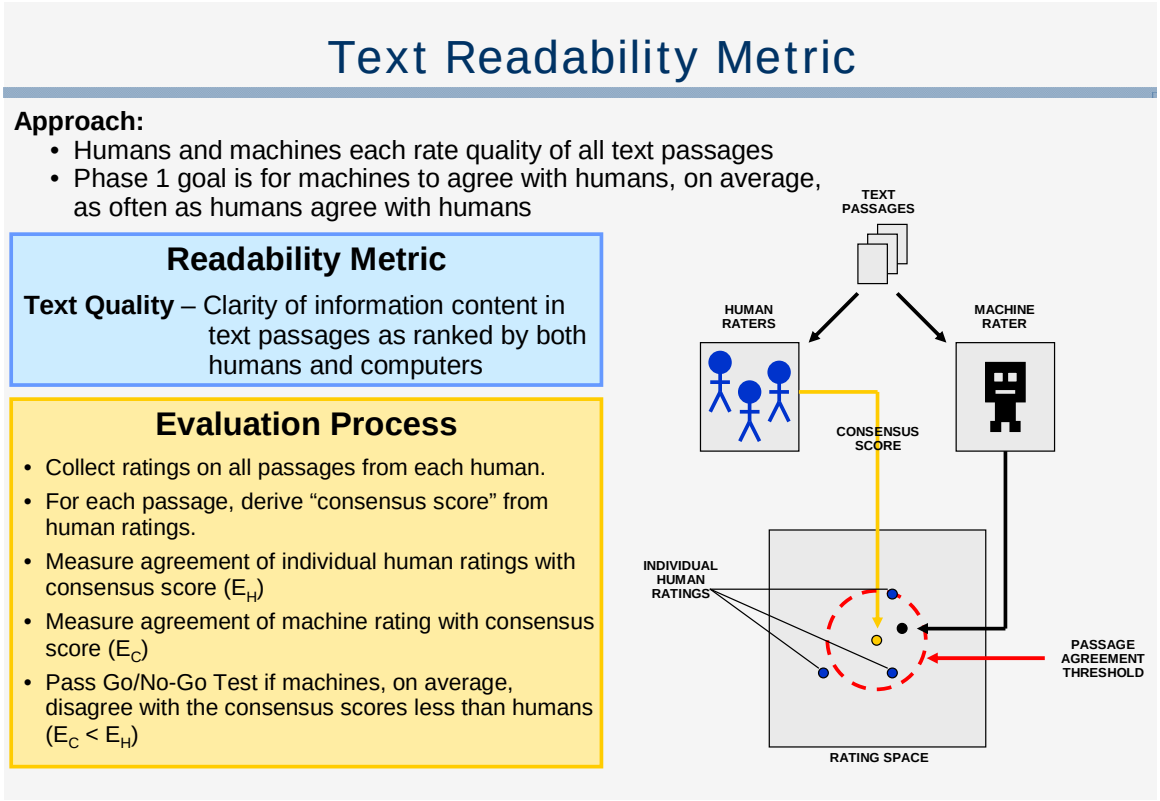

Figure 5 - Text Readability Metric Description

The Reading Teams will be provided with training data consisting of sample passages and associated human rating data. This will be provided to the Reading Teams relatively early in Phase 1 by the Evaluation Team.

During the end-of-phase evaluations for Phase 1, the following process will be used to evaluate the Reading Systems against the Text Readability Metric:

Start with H expert human raters and
M passages.

Let $R_{h,m}$ be the rating by human expert *h* of passage *m*,
$R_{C,m}$ be the rating by the computer of passage *m*, and
$d(R_{h,m},R_{h',m})$ be the distance between two ratings of a passage

Humans generate ratings for each passage

Compute consensus score for each passage, $R_{G,m}$, that minimizes its distance from human ratings: $R_{G,m} = \arg\min_{R'}\left(\sum_{h=1}^{H} d\left(R', R_{h,m}\right)\right)$

Calculate average disagreement between humans and consensus score:

$$E_{H,m} = \frac{1}{H}\sum_{h=1}^{H} d\left(R_{h,m}, R_{G,m}\right)$$

Reading systems generate ratings for each passage

Calculate disagreement between computer and consensus score:

$$E_{C,m} = d\left(R_{C,m}, R_{G,m}\right)$$

Calculate computer and average human disagreement across all passages:

$$E_H = \frac{1}{M}\sum_{m=1}^{M} E_{H,m} \quad , \quad E_C = \frac{1}{M}\sum_{m=1}^{M} E_{C,m}$$

Phase 1 Go/No-Go Criteria: $E_C \leq E_H$

I.C.1.c.ii Correctness/Completeness

The performance of the Reading Systems, along the dimensions of correctness and completeness in answering queries, will be compared to that of humans answering the same queries. Operationally the Reading Systems are expected to be used in executing performance tasks, but the lower-level interfaces will be at the query-response level. Humans will be given the queries in natural language while the Reading

Systems will be given queries specified in a formal language. The F-measures (the – possibly weighted – harmonic mean of precision and recall) of Reading System performance and human performance will be compared. The ratio of machine to human performance is expected to increase every year.

Two different, but closely related, performance ratios (Figure 6) will be used during the program. Answering "Fact"-based questions from text is tracked and measured separately from Q&A that requires inference over multiple facts and rules. Thus the first of these metrics, Performance Ratio for Facts (PR_F), measures performance against queries for which ground facts can be found explicitly stated in the source text. The second, Performance Ratio for Inference (PR_I), measures performance against queries for which some level of logic and inference is required in order to find the answers. We define a factual query to be a query that is answered based on an assertion drawn from a single sentence in the text. Inferential queries must be computed by an inference chain containing more than one assertion typically drawn from text in multiple sentences in the text.

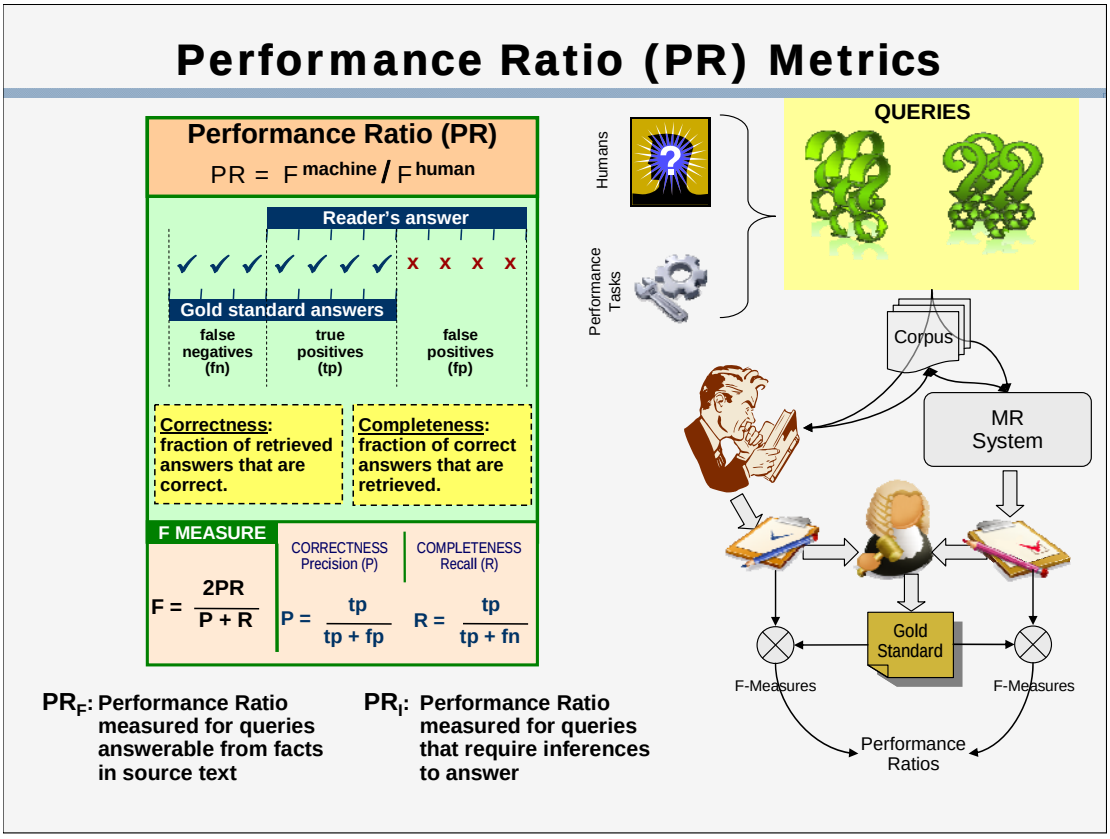

Figure 6- Performance Ratio Metrics Description

The Evaluation Team will determine the "consensus" result set for selected task-generated queries. Queries will be based on an underlying AI performance task, although they may not be literal queries taken from this task. Both the correctness and completeness of the Reading Systems' responses will be assessed relative to these

human-generated consensus answers. Systems will be assessed against humans, rather than against the underlying performance task, for many reasons. Performance on the underlying performance task could succeed or fail based on a number of factors. By comparing against human consensus answers, the performance of the systems in bridging from natural texts to formal language is isolated and can be assessed. Since the program will use natural texts, no guarantees are practical for any specific levels of completeness or correctness. By comparing the systems against consensus responses, assessments will be against the fraction of achievable responses obtained, rather than the fraction of queries answered.

Employing consensus reports in the assessment mechanism also enables control for the complexity of the reading task undertaken. The source texts cannot be easily constrained without introducing distortions into the properties of the text. However, selective enforcement of the types of textual inferences considered to be “in bounds” provides an important control for focusing the efforts of the program. For example, at the beginning of the program the focus is on the capture of ground assertions that do not require inference from multiple sentences. The assessment can be focused accordingly by instructing judges to exclude any assertion that they cannot connect to a single sentence in the corpus. In later phases of the program, when inference is expected and must be measured, queries can be categorized according to whether or not they can be obtained *without* inference. Over the course of the program, the selection of queries and responses used in assessments will be varied to align assessments with they focus of the program from phase to phase.

During the end-of-phase evaluations, the following process will be used to evaluate the Reading Systems against the relevant Performance Ratio metrics:

- Query Creation. For each domain corpus, the Evaluation Team generates a set of questions that are to be answered from the corpus. These queries may be extracted from underlying queries executed as part of a performance task. In all cases the queries will be formulated both as natural language questions for humans as well as formally specified queries for submission to the Reading Systems.
    - A set of questions for which answers can be found in that corpus as explicitly-stated facts or assertions will be generated for use in computing PR_F;
    - A set of questions for which answers *cannot* be found in the domain corpus as explicitly-stated facts or assertions but which can be answered using inferences appropriate to the current phase of the program will be generated for use in computing PR_I.
- Humans and Reading Systems **generate result sets** of answers with provenance from the source text for each answer.
- The union of these sets is vetted by human judges to determine which answers are correct – these become the gold standard answer set for that query.
- Calculate F-measures. Precision is the fraction of answers found by a subject that are in the gold set; recall is the fraction of gold set answers in all of the answers found by the subject.
- Calculate Performance Ratio as percentage of human performance $PR=100*(F_{machine}/F_{human})$

I.C.1.c.iii Effort Reduction

Some approaches to machine reading may require nominal amounts of knowledge engineering for each new reading domain. This knowledge engineering time is expected to be a fraction of the time it would take to manually read and encode the knowledge in the corpus. The Effort Reduction metric (Figure 7) is the ratio of the (calculated) effort required to populate a knowledge base using traditional means to the effort required to customize the Reading System to function in a new domain.

The goal of the MRP is to develop general purpose technology for employing natural text in formal contexts. Notionally, this would be a “magic box” that automatically transforms natural text into formal contexts. Practically, there are limits to the degree of generality and automation that is possible. Acknowledging this, the program accepts the controlled use of task-specific engineering in the MRP.

The Effort Reduction metric will compare how much effort is expended tweaking a Reading System to accommodate a new task domain with how much manual knowledge engineering would have been required to manually encode the knowledge that a Reading System ultimately captures from a supplied corpus. Beginning in Phase 3 of the MRP, there will be task domains used for assessment that are hidden from the Reading Teams until the assessment period. For these hidden domains, the amount of

engineering required to *adapt* the Reading System to the new performance task and domain type will be compared to (an approximation of) the effort require to manually construct a knowledge base appropriate for the hidden domain. Since this solicitation encourages a broad range of technical approaches, the nature of the specialization effort is not restricted here. Nonetheless, the general applicability of reading technology should be considered in the adaptation approach. Examples include providing specialized forms of training data or manually tweaking parameters on the algorithms in order to optimize performance in the target. The intent is simply to quantitatively compare the cost of tweaking a Reading System to a domain to the cost of manually constructing a knowledge base for a domain.

It is not practical or cost effective to manually create complete knowledge bases that parallel those built by the Reading Systems, just to have a numerator for this metric. Instead, knowledge engineers on the Evaluation Team will be tasked to spend a bounded amount of time extracting facts and rules from the source text, using state of the art tools, and the output of this fixed duration effort will be extrapolated linearly to calculate how much time would have been required had the effort continued to build a complete knowledge base. This extrapolated time will be compared to the amount of time required to adapt a Reading System to transform text to knowledge in the same domain in order to produce the effort reduction ratio for that task. Figure 7 depicts this calculation on the left. The temporal order of information flows between the teams during the calculation of effort reduction is shown on the right.

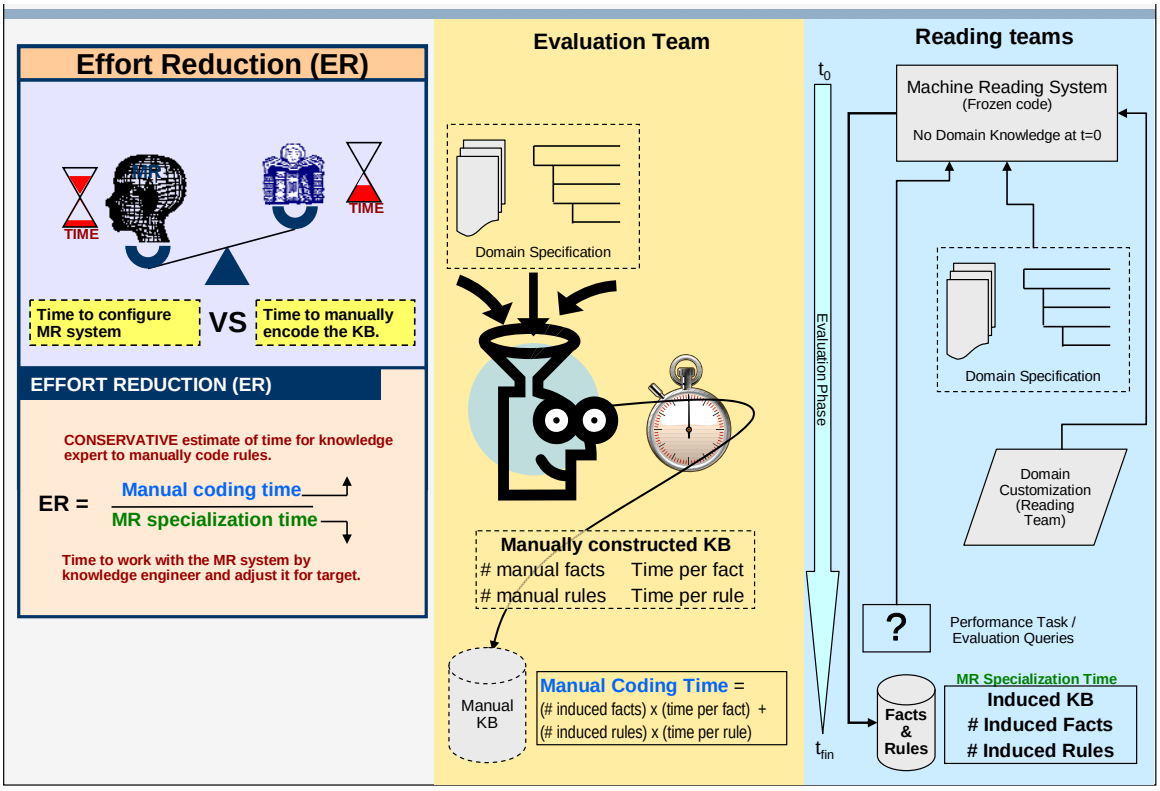

Figure 7 - Effort Reduction Metric

I.C.1.d. Go/No-Go Targets by Program Phase

The performance targets that must be achieved by the Reading Systems for each phase of the program are shown in Figure 8. A given metric applies only in those phases for which a value for that metric is shown (e.g., Effort Reduction is not a Go/No-Go criterion until Phase 3).

	METRIC	1	2	3	4	5
Correctness/ Completeness	Text Readability	100% of Humans				
	Performance Ratio$_{Facts}$		40	50	70	85
	Performance Ratio$_{Inference}$		10	15	35	60
	Effort Reduction			10 x	100x	**1000x**

Figure 8 - Go/No-Go Thresholds by Phase

I.C.2. Evaluation Team

This section provides a summary of the tasks that the Evaluation Team will be expected to perform. Terminology not defined in this section is defined earlier in the solicitation and/or in the Terminology Appendix (VIII.B.1.).

I.C.2.a. Task Details

Expectations over All Phases

- Work with DARPA and other government organizations to:
  - Ensure the evaluation metrics can be clearly applied during each phase and to clarify their definition and application when needed;
  - Ensure the evaluation process is fair and appropriate to the current phase and that evaluation queries exercise the appropriate capabilities of the Reading Systems; and
  - Conduct end-of-phase evaluations of Reading Team(s) and prepare reports for DARPA reporting the results of the evaluations.
- Ensure that domain corpora developed each year are appropriate and adequate for evaluating the capabilities of the Reading Systems. This includes the style and complexity of the writing, as well as the amount of domain-relevant information contained in the text that the Reading Systems are expected to accommodate.
- Conduct end-of-phase Go/No-Go evaluations of Reading Systems.
- Support DARPA in developing briefing materials for annual Program Status briefing and Go/No-Go briefing.
- Deliver reading tasks to the Government and the Reading Teams. Domain selections and corpus sizes for reading tasks shall be approved by DARPA. Reading tasks shall be appropriate, in size and content, to the MRP's expectations

for the current program phase. At least one new domain per phase shall exercise the Reading Systems on each of the qualitative challenge dimensions of the MRP for that phase.
  - As described in the Program Description section, reading tasks will have an associated corpus, one or more Syntax Specifications and Mappings with associated queries, and consensus responses.
  - One new reading task domain will be produced and delivered to the Reading Teams at the midpoint of each phase. Deliver hidden reading tasks to Reading Teams in time to carry out the end-of-phase evaluation including the effort reduction engineering that will occur beginning in Phase 3.
- In collaboration with DARPA and the Reading Teams, develop, specify, and revise:
  - Query interfaces and response formats;
  - The inference interface, which provides access to the DSRSs; and
  - Common interfaces for components found to be common across the reading architectures to facilitate decomposing the MR process into a small number of key subtasks.
- Support the creation of an infrastructure to nurture a larger machine reading research community. Toward this goal,
  - Identify, capture, organize and maintain key components in Reading Systems that can be adapted and shared among many Reading Systems.
  - Acquire and/or develop and curate training and testing datasets
- Support transition activities associated with the MRP. This includes discussions with potential transition partners as well as shaping the reading tasks to align with transition opportunities.

Expectations for Phase 1

- Assist the Government in planning for future machine reading evaluations that will be conducted independently from the MRP and will be open to all interested parties.
- Reading Tasks
  - Develop two complete reading tasks, with corpora on the order of a hundred pages, which contain a significant amount of text that is written in a clear, unambiguous style and includes factual information written in styles compatible with the minimum linguistic capabilities required of the Reading Teams for Phases 1 and 2.
  - Assemble a test corpus sufficient for evaluating the performance of the Reading Teams against the Text Readability Metric.
    - Provide human ratings, as to the clarity and ambiguity of the writing employed, of documents in this corpus.
    - Provide Reading Teams with training data for their evaluation against the Text Readability Metric. This will include sample passages and ratings of these passages made by human reviewers.

Expectations for Phases 2-5

- Deliver two new reading task domains each phase to the Government and the Reading Teams. Beginning in Phase 3, at least one of the new reading task

domains developed in each phase is to be kept hidden from the Reading Teams. No information regarding the hidden domains may be delivered to the Reading Teams prior to the beginning of Go/No-Go evaluations for each phase.
  - Task Corpora
    - Phase 2: At least one of the new task corpora will be on the order of a thousand pages of text.
    - Phases 3 and 4: At least one of the new task corpora will be on the order of 3x to 5x the size of the largest corpus from the previous phase
    - Phase 5: At least one of the new task corpora will be "web scale" (e.g., 10x to 100x larger than any previous corpus).
  - Provide a DSRS for each new task domain.
  - Develop Evaluation Queries and consensus Answers for new domains in time for the end-of-phase Go/No-Go evaluations.
  - Deliver one reading task domain (with all associated components) at the midpoint of each phase.
- Maintain reading tasks developed during earlier program phases, as directed by DARPA, and update as required to remain consistent with any changes made to interfaces and specifications. This role should be small relative to the costs of delivering new reading task domains. DARPA may agree, in negotiation with the Evaluation Team, to terminate maintenance of some prior domains if the maintenance costs become excessive.
- Conduct end-of-phase evaluations of Reading Teams.
  - Manually construct a knowledge-base subset for each new hidden domain (beginning in Phase 3) to use effort required as the comparison baseline for evaluating the performance of the Reading Systems against the Effort Reduction metric.
- Assist the Government in conducting open evaluation of Reading Systems and updating and revising future plans with lessons learned.

I.C.2.b. Performance Expectations and Deliverables

Some of the key artifacts that the Evaluation Team is responsible for producing and maintaining are described here. Greater detail (including information on subcomponents of the various artifacts) is found in the Terminology Appendix (VIII.B.1.)

Reading Task Domain Specification - A formal specification of the data and control between the domain-specific tasks (supplied by the Evaluation Team) and the Reading Systems (supplied by the Reading Teams). This interface is specified in this document, but is expected to be managed and evolved under the Government's guidance by the Evaluation Team as program needs dictate.

Reading Task Domains - The set of components that collectively define a reading and reasoning context. This includes the testing and background corpora, one or more Syntax Specifications and Mappings, a DSRS for accessing task-specific background knowledge, and a set of performance-task queries with consensus results.

- **Testing Corpus** - The set of texts that contain information required to answer the task queries.
- **Background Corpus** - Optional texts that may be supplied by the Evaluation Team for some task domains.
- **Syntax Specification and Mapping** - Specification of the syntactic structure of the formal knowledge used in the target AI task (Syntax Specification) and a set of example pairs linking formal assertions to natural text sentences and phrases with the same meaning as those formal assertions (Mapping).
- **Domain-Specific Reasoning System** - A component that is capable of performing inferences relevant to the associated reading task domain. The DSRS must be compliant with the Inference Interface, but the details of how it operates is irrelevant (i.e., it is treated by the MRP as a "black box"). This is also the vehicle by which reading systems access task-specific background knowledge. The DSRS is queried using the inference interface.
    - **Inference Interface** - The interface used by the Reading System to query the DSRS. The basic interface is expressed as probabilistic entailment. That is, given an assertion set and a query template, the inference interface returns a result set specifying assertions that follow from the assertions set, and available background knowledge that also match the query template.
    - **Background Knowledge** - Task-specific knowledge that the Evaluation Team and DARPA deem necessary for the reading and reasoning task and that is made available via the DSRS.
- **Performance Task Queries and Consensus Responses** - Task-specific reasoning queries generated (or motivated) by a domain-specific performance task (Performance Task Queries). Since Go/No-Go metrics are based on comparisons to human performance on the same task, a consensus set of human response assertions to the queries will also be produced (Consensus Responses).

I.C.2.c. Metrics (Go/No-Go Targets)

The Go/No-Go criterion for the Evaluation Team is meeting the deliverables for each program phase. Achieving the Go/No-Go for each phase of the program is a necessary condition for establishing eligibility for continuing to work on subsequent phases of the MRP.

II. <u>AWARD INFORMATION</u>

Multiple awards are anticipated for the Reading Teams and a single award is anticipated for the Evaluation Team.

The amount of resources made available to this BAA will depend on the quality of the proposals received and the availability of funds. The Government reserves the right to select for negotiation all, some, one, or none of the proposals received in response to this solicitation, and to make awards without discussions with offerors. The

Government also reserves the right to conduct discussions if the Source Selection Authority later determines them to be necessary. If warranted, portions of resulting awards may be segregated into pre-priced options. Additionally, DARPA reserves the right to accept proposals in their entirety or to select only portions of proposals for award. In the event that DARPA desires to award only portions of a proposal, negotiations may be opened with that offeror. If the proposed effort is inherently divisible and nothing is gained from the aggregation, offerors should consider submitting it as multiple independent efforts. The Government reserves the right to fund proposals in phases, with options for continued work at the end of one or more of the phases.

Awards under this BAA will be made to offerors on the basis of the evaluation criteria listed below (see section V - Application Review Information) and program balance to provide best value to the Government. Proposals identified for negotiation may result in a contract, grant or cooperative agreement depending upon the nature of the work proposed, the required degree of interaction between parties, and other factors. The Government reserves the right to request any additional, necessary documentation once it makes the award instrument determination. Such additional information may include but is not limited to Representations and Certifications.

As of the date of publication of this BAA, DARPA expects that program goals for this BAA may be met by offerors intending to perform 'fundamental research,' i.e., basic and applied research in science and engineering, the results of which ordinarily are published and shared broadly within the scientific community, as distinguished from proprietary research and from industrial development, design, production, and product utilization the results of which ordinarily are restricted for proprietary or national security reasons. Notwithstanding this statement of expectation, DARPA is not prohibited from considering and selecting research proposals that, while perhaps not qualifying as “fundamental research” under the foregoing definition, still meet the BAA criteria for submissions. In all cases, the contracting officer shall have sole discretion to select award instrument type and to negotiate all instrument provisions with selectees.

III. ELIGIBILITY INFORMATION

III.A. Eligible Applicants

All responsible sources capable of satisfying the Government's needs may submit a proposal that shall be considered by DARPA. Historically Black Colleges and Universities (HBCUs), Small Businesses, Small Disadvantaged Businesses and Minority Institutions (MIs) are encouraged to submit proposals and join others in submitting proposals. However, no portion of this announcement will be set aside for Small Business, Small Disadvantaged Business, HBCU and MI participation due to the impracticality of reserving discrete or severable areas of this research for exclusive competition among these entities.

Federally Funded Research and Development Centers (FFRDCs) and Government entities (Government/National laboratories, military educational institutions, etc.) are

subject to applicable direct competition limitations and cannot propose to this BAA in any capacity, unless they can clearly demonstrate the work is not otherwise available from the private sector AND they also provide written documentation citing the specific statutory authority (as well as, where relevant, contractual authority) establishing their eligibility to propose to government solicitations. At the present time, DARPA does not consider 15 U.S.C. 3710a to be sufficient legal authority to show eligibility. While 10 U.S.C. 2539b may be the appropriate statutory starting point for some entities, specific supporting regulatory guidance, together with evidence of agency approval, will still be required to fully establish eligibility. DARPA will consider eligibility submissions on a case-by-case basis; **however, the burden to prove eligibility for all team members rests solely with the Offeror.**

Foreign entities and individuals may participate to the extent that such participants comply with any necessary Non-Disclosure Agreements, Security Regulations, Export Laws, and other governing statutes and regulations applicable under the circumstances.

Applicants considering classified submissions (or requiring access to classified information during the life-cycle of the program) shall ensure all industrial, personnel, and information system processing security requirements are in place and at the appropriate level (e.g., Facility Clearance (FCL), Personnel Security Clearance (PCL), certification and accreditation (C&A)) and any Foreign Ownership Control and Influence (FOCI) issues are mitigated prior to such submission or access. Additional information on these subjects can be found at: www.dss.mil.

III.A.1. Procurement Integrity, Standards of Conduct, Ethical Considerations, and Organizational Conflicts of Interest

Current federal employees are prohibited from participating in particular matters involving conflicting financial, employment, and representational interests (18 USC 203, 205, and 208.). The DARPA Program Manager for this BAA is Dr. Daniel Oblinger. As of the date of first publication of the BAA, the Government has not identified any potential conflicts of interest involving this Program Manager. Once the proposals have been received, and prior to the start of proposal evaluations, the Government will assess potential conflicts of interest and will promptly notify the offeror if any appear to exist. (Please note the Government assessment does NOT affect, offset, or mitigate the offeror's own duty to give full notice and planned mitigation for all potential organizational conflicts, as discussed below.) The Program Manager is required to review and evaluate all proposals received under this BAA and to manage all selected efforts. Offerors should carefully consider the composition of their performer team before submitting a proposal to this BAA.

In accordance with FAR 9.503 and without prior approval or a waiver from the DARPA Director, a contractor cannot simultaneously be a SETA and a performer. Therefore, all offerors and proposed subcontractors must affirm whether they (their organizations and individual team members) are providing scientific, engineering, and technical assistance

(SETA) or similar support to any DARPA technical office(s) through an active contract or subcontract. All affirmations must state which office(s) the offeror, sub and/or individual supports and identify the prime contract numbers. Affirmations shall be furnished at the time of proposal submission. All facts relevant to the existence or potential existence of organizational conflicts of interest (FAR 9.5) must be disclosed. The Government will make the final determination on what constitutes a conflict of interest. The disclosure shall include a description of the action the offeror has taken or proposes to take to avoid, neutralize, or mitigate such conflict. **Proposals that fail to fully disclose potential conflicts of interests and/or do not have plans to mitigate this conflict will be rejected without technical evaluation and withdrawn from further consideration for award.**

If a prospective offeror has any questions on what constitutes a conflict of interest (whether organizational or otherwise), the offeror should promptly raise the issue with DARPA by sending his/her contact information and a summary of the potential conflict by email to the mailbox address for this BAA at DARPA-BAA-09-03@darpa.mil, before time and effort are expended in preparing a proposal and mitigation plan. If, in the sole opinion of the Government after full consideration of the circumstances, any conflict situation cannot be effectively mitigated, the proposal may be rejected without technical evaluation and withdrawn from further consideration for award under this BAA.

III.B. Cost Sharing or Matching

Cost sharing is not required for this particular program; however, cost sharing will be carefully considered where there is an applicable statutory condition relating to the selected funding instrument (e.g., for any Technology Investment Agreement under the authority of 10 U.S.C. 2371).

III.C. Other Eligibility Requirements

A performer selected for the Evaluation effort, will not and cannot be selected for *any* portion of the Reading Team effort, whether as a prime or subcontractor or in any other capacity; therefore, if DARPA selects your proposal for the Evaluation Team, your proposal submitted for the Reading Team will be considered as “not selectable” even if it would otherwise have been considered “selectable” according to the evaluation criteria. This is to avoid organizational conflict of interest situations between technical and evaluation efforts and to ensure objective test and evaluation results. The Government reserves the right to choose which task proposal to select and which not to select, in cases where an offeror has submitted otherwise selectable proposals to both tasks.

IV. APPLICATION AND SUBMISSION INFORMATION

IV.A. Address to Request Application Package

This solicitation contains all information required to submit a proposal. No additional forms, kits, or other materials are needed. This notice constitutes the total BAA. No additional information is available, nor will a formal Request for Proposal (RFP) or additional solicitation regarding this announcement be issued. Requests for same will be disregarded.

IV.B. Content and Form of Application Submission

All proposals must be zipped and encrypted using Winzip or PKZip with 256-bit AES encryption. Only one zipped/encrypted file will be accepted per proposal. Proposals which are not zipped/encrypted will be rejected by DARPA. An encryption password form must be completed and emailed to DARPA-BAA-09-03@darpa.mil at the time of proposal submission. See https://www.CSC-Ballston.com/baa/Encryption_Instructions.htm for the encryption password form and additional encryption information. Note: the word "PASSWORD" must appear in the subject line of the above email and there are minimum security requirements for establishing the encryption password. Failure to provide the encryption password will result in the proposal not being evaluated.

IV.B.1. Proposal Information

DARPA will employ an electronic upload submission system for all responses to this BAA. Responding to this announcement requires completion of an online cover sheet for each proposal prior to submission. To do so, the offeror must go to https://csc-ballston.com/baa/index.asp?BAAid=09-03 and follow the instructions there. Upon completion of the online cover sheet, a Confirmation Sheet will appear along with instructions on uploading proposals. The Confirmation Sheet will be used as the Cover Sheet for the proposal and will contain the information outlined below in Proposal Section 1.1. If an offeror intends to submit more than one proposal, a unique UserId and password must be used in creating each cover sheet. **Since offerors may encounter heavy traffic on the web server, they SHOULD NOT wait until the day the proposal is due to fill out a coversheet and submit the proposal!**

IV.B.2. Proposal Preparation and Format

The proposal shall be delivered in two volumes, Volume 1 (technical proposal) and Volume 2 (cost proposal). Proposals not meeting the format described in this BAA may not be reviewed.

Volume 1 – Technical Proposal

The technical proposal shall include the following sections, each starting on a new page (where a "page" is 8-1/2 by 11 inches with type not smaller than 12 point, margins not smaller than 1 inch, and line spacing not smaller than single-spaced). All submissions must be in English. Individual elements of the proposal shall not exceed the total of the maximum page lengths for each section as shown in braces { } below.

Ensure that each section provides the detailed discussion of the proposed work necessary to enable an in-depth review of the specific technical and managerial issues. Specific attention must be given to addressing both risk and payoff of the proposed work that make it desirable to DARPA.

Proposal Section 1 -- Administrative and Technical Details

1.1 Confirmation Sheet/Cover Sheet

As described above, this cover sheet will contain the following information:

- BAA number;
- Proposal title;
- Reading Team or Evaluation Team
- Technical point of contact including: name, telephone number, electronic mail address, fax (if available) and mailing address;
- Administrative point of contact including: name, telephone number, electronic mail address, fax (if available) and mailing address;
- Summary of the costs of the proposed research, including total base cost, estimates of base cost in each year of the effort, estimates of itemized options in each year of the effort, and cost sharing if relevant;
- Contractor's reference number (if any)
- Contractor's type of business, selected from among the following categories:
    - WOMEN-OWNED LARGE BUSINESS,
    - OTHER LARGE BUSINESS,
    - SMALL DISADVANTAGED BUSINESS [Identify ethnic group from among the following: Asian-Indian American, Asian-Pacific American, Black American, Hispanic American, Native American, or Other],
    - WOMEN-OWNED SMALL BUSINESS,
    - OTHER SMALL BUSINESS,
    - HBCU,
    - MI,
    - OTHER EDUCATIONAL,
    - OTHER NONPROFIT, OR
    - FOREIGN CONCERN/ENTITY.

1.2 Table of Contents {No page limit}

1.3 PowerPoint (Version MS Office 2003) Summary Overview Slide {1 chart}
Provide a one slide summary of the proposal in PowerPoint that effectively and succinctly conveys the main objective, key innovations, expected impact, and the ***truly unique*** aspects of the technology you are planning to develop to solve the objectives of the program. Inclusion of both a summary graphic and verbiage is required with compelling full story differentiating the offeror's approach. It is anticipated this slide will contain redundant information contained elsewhere in the proposal. Approximately ~100 words in the Notes section required to support the slide. **Note: This slide must be severable from the rest of the proposal with all its PowerPoint functionality.**

1.4 Power Point (MS Version 2003) Justification Slide {1 chart}

This chart justifies team size and all activities of team members. This slide must reflect the decomposition of the set of activities which are comprised to accomplish the overall program objectives. Within this decomposition, each set of activities must clearly define the group doing the proposed work (prime and any sub contractors). Costing estimates should be provided for any/all activities anticipated in Phase 1 and should total the budget for Phase 1. Include approximately 100 words in the Notes section to support the slide. **Note: This slide must be severable from the rest of the proposal with all its PowerPoint functionality.**

1.5 Proposal Overview {1 Page}

Provides big picture summary of the proposal.

1.6 Innovative Claims for the Proposed Research {1 Page}

This page is the centerpiece of the proposal and should succinctly describe the proposed unique approach and concepts using a bullet format in this form:

- "**Name of concept**" - *This is a short bullet phrase describing the important concept.* (Please re-use the "name of concept" in the Technical Approach section and other parts of the proposal as needed.)

1.7 Proposal Roadmap {1 Page}

The roadmap provides a top-level view of the content and structure of the proposal. It contains a synopsis for each of the roadmap areas defined below, which should be elaborated elsewhere. It is important to make the synopses as explicit and informative as possible. The roadmap must also cross-reference the proposal page number(s) where each area is elaborated. The required roadmap areas are:

a. Main goals of the proposed research.
b. Tangible benefits to end users (i.e., benefits of the capabilities afforded if the proposed technology is successful).
c. Critical technical barriers (i.e., technical limitations that have, in the past, prevented achieving the proposed results).
d. Main elements of the proposed technical approach.
e. Basis of confidence (i.e., rationale that builds confidence that the proposed approach will overcome the technical barriers).
f. Nature and description of end results to be delivered to DARPA. In what form will results be developed and delivered to DARPA and the scientific community? Note that DARPA encourages experiments, simulations, specifications, proofs, etc. to be documented and published to promote progress in the field. Offerors should specify both final and intermediate products.
g. Cost and schedule of the proposed effort.

1.8 Technical Approach {15 pages}

Provide a detailed description of the technical approach for the complete program. Phase 1 must be addressed in a separate subsection from the rest of the phases.

There are no restrictions on page allocations between subsections. This section will elaborate on many of the topics identified in the proposal roadmap and will serve as the primary expression of the offerors' scientific and technical ideas. A one paragraph executive-style summary must be provided for each proposal subsection. The following sections must be included within this portion of the proposal but is not limited to these items:

a. Problem Description. Provide a concise description of the problem areas addressed specific to your approach.
b. Research goals. Identify specific research goals. Goals should address the technical challenges of the effort.
c. Expected impact. Describe the expected impact of your research
d. Approach to be taken in the proposal.
e. Justify ability to pass Go/No-Go evaluation criteria requirements for specified metrics for each phase. Possibilities might include:
    - A summary to enumerate, prioritize, and quantify key risks
    - Approaches to mitigate these risks
    - Checkpoint milestones to assess progress
    - Other monitoring or programmatic solutions

1.9 Comparison with Current Technology {2 Pages}

Describe state of the art approaches and the limitations that relate to each area addressed by the proposal. Describe and analyze state of the art results, approaches, and limitations within the context of the problem area addressed by this research. Demonstrating problem understanding requires not just the enumeration of related efforts; rather, related work must be compared and contrasted to the proposed approach.

1.10 Statement of Work (SOW) {5 pages}

In plain English, clearly define the technical tasks/subtasks to be performed, their durations, and dependencies among them. For each task/subtask, provide:

- A general description of the objective (for each defined task/activity);
- A detailed description of the approach to be taken to accomplish each defined task/activity);
- Identification of the primary organization responsible for task execution (prime, sub, team member, by name, etc.);
- The exit criteria for each task/activity - a product, event or milestone that defines its completion.
- Define all deliverables (reporting, data, reports, software, etc.) to be provided to the Government in support of the proposed research tasks/activities.

Note: The SOW should be developed so that each phase of the program is separately defined. Offerors should format their proposals for Phase 1 with subsequent phases (e.g., Phases 2, 3, 4, 5) tasks/subtasks as options. Do not include any proprietary information in the SOW.

1.11 Deliverables Description {3 Pages}

List and provide by phase a detailed description for each proposed deliverable, including receiving organization and expected delivery date for each deliverable. Include in this section all proprietary claims to results, prototypes, or systems supporting and/or necessary for the use of the research, results, and/or prototype. If there are no proprietary claims, this should be stated. The offeror must submit a separate list of all technical data or computer software that will be furnished to the Government with other than unlimited rights. See section VI.B.2 below for more information

1.12 Management Plan {3 Pages}

Describe formal teaming agreements that are required to execute this program, a brief synopsis of all key personnel, and a clearly defined organization chart for the program team (prime contractor and subcontractors, if any, with location addresses). Provide an argument that the team size and composition are both necessary and sufficient to meet the program objectives. Provide task descriptions, costs, and interdependencies for each individual effort and/or subcontractor. To the extent that graduate students and postdoctoral fellows are involved in individual efforts, describe their role and contribution. Information in this section must cover the following information:

a. Programmatic relationship of team members;
b. Unique capabilities of team members;
c. Task responsibilities of team members;
d. Teaming strategy among the team members;
e. Key personnel; and
f. Government role in project, if any.

The purpose of the Management Plan is to convey the roles and responsibilities for team members and the overarching plan for managing and evolving the team as the MRP progresses. Where the information specified in this section (e.g., key personnel, task descriptions) is also required in subsequent sections, offerors should focus on how that information is germane to the program management and defer details to the subsequent sections.

1.13 Schedule and Milestones

This section should include:

a. {1 Page} Schedule Graphic. Provide a graphic representation of project schedule including detail down to the individual effort level. This should include but not be limited to, a multi-phase development plan, which demonstrates a clear understanding of the proposed research; and a plan for periodic and increasingly robust tests over the project life that will show applicability to the overall program concept. Show all project milestones. Use “x months after contract award” designations for all dates.
b. {3 Pages} Detailed Task Descriptions. Provide detailed task descriptions for each discrete work effort and/or subcontractor in schedule graphic.
c. {3 Pages} Project Management and Interaction Plan. Describe the project management and interaction plans for the proposed work. If proposal includes subcontractors that are geographically distributed, clearly specify working/meeting models. Items to include in this category include software/code

repositories, physical and virtual meeting plans, and online communication systems that may be used.

1.14 Personnel, Qualifications, and Commitments {NO MORE THAN ONE PAGE PER KEY PERSON}

List key personnel, showing a concise summary of their qualifications, discussion of offeror's previous accomplishments, and work in this or closely related research areas. Indicate the level of effort in terms of hours to be expended by each person during each contract year and other (current and proposed) major sources of support for them and/or commitments of their efforts. DARPA expects all key personnel associated with a proposal to make substantial time commitment to the proposed activity and the proposal will be evaluated accordingly. It is DARPA's intention to incorporate key personnel clauses into any resultant award(s); therefore, offerors should ONLY propose personnel whom they intend to have perform.

Include a table of key individual time commitments as follows:

Key Individual	**Project**	**Pending/Current**	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**	**Phase 5**
Jane Doe	Machine Reading	Proposed	ZZZ hours (%)	UUU hours (%)	WWW hours (%)		
	Project 1	Current	n/a	n/a	n/a		
	Project 2	Pending	100 hours (%)	n/a	n/a		
John Deer	Machine Reading	Proposed					

1.15 Cost Summaries {2 pages}

This section shall contain two tables: the first table must summarize the proposed costs but break them down by project task, subtask, and phase, i.e., show the costs of each project task and subtask for each phase, by month, with the task and subtask labels on the y-axis and the program phases on the x-axis. It may be appropriate to create a subtotal under some closely related tasks. Table entries should contain the dollar figure and a percentage that specifies the percentage of that phase's total costs that are allocated to said task.

The second table should show the costs broken down by prime/subcontractor by month, by phase, i.e., the labels of the prime/subcontractors should be on the y-axis and the three phases on the x-axis. Table entries should contain the dollar figure and a percentage that specifies the percentage of that phase's total costs allocated to said prime or subcontractor. Offerors should format their proposals for Phase 1, with subsequent phases (e.g., Phase 2, 3, 4, 5) priced as options.

1.16 Organizational Conflict of Interest Affirmations and Disclosure {No page limit}
Per the instructions in Section III.C.1 above, all offerors and proposed subcontractors must provide documentation showing whether they (their organizations and individual team members) are providing scientific, engineering, and technical assistance (SETA) or similar support to any DARPA technical office(s) through an active contract or subcontract. All affirmations must state which office(s) the offeror, sub and/or individual supports and identify the prime contract numbers. If the offeror or any proposed sub IS providing SETA support as described (regardless of which DARPA technical office is being supported), then the offeror shall include a description of the action the offeror has taken or proposes to take to avoid, neutralize, or mitigate such conflict. The Government will make the determination of what constitutes a conflict of interest. If the offeror or any proposed sub IS NOT currently providing SETA support as described, then the offeror should simply state "NONE."

Proposals that fail to fully disclose potential conflicts of interests or do not have acceptable plans to mitigate identified conflicts will be rejected without technical evaluation and withdrawn from further consideration for award.

1.17 Intellectual Property {No page limit}
Per section VI.B.2 below, offerors responding to this BAA shall identify any intellectual property restrictions. If no restrictions are intended, then the offeror should state "NONE".

1.18 Licensing Model {5 Pages}
In addition to the conventional licensing in section VI.B.2 below, this program seeks to extend and distribute materials developed in this program in order to further the science of machine reading. Therefore, this section has been included to request proposals for specific broader licensing terms that would be acceptable to the performers. Please document clearly the licensing model, explain how it will contribute to the licensing goals of the program as stated below, identify any costs that the government would incur which are associated with licensing requirements, and provide one or more sample licensing agreements.

<u>Goals</u>

A central goal of the Machine Reading Program is to foster a research community in machine reading. As of the drafting of this solicitation, it is unclear whether or not the public release of materials developed in the course of this research program would support this goal. Furthermore, if components were to be released, it isn't clear which would best serve these interests. Thus, DARPA intends to work closely with the program performers and seek their advisement on these issues prior to setting public release requirements.

However, in anticipation of the potential public release of program materials, licenses should be constructed to minimize barriers. In particular, where applicable, the following primary use cases should be addressed:

- Ongoing extension and reuse of training and evaluation data sets
- Ongoing extension and reuse of the entire delivered Reading System
- Ongoing extension and reuse of individual machine reading algorithm components

Guidelines

The following use restriction scale is based upon divisions that are considered relevant to this program. Therefore, this constitutes an appropriate metric for use in the section of the proposal that describes the license(s), and will also be used by the Government when evaluating your plan. Please note that this scale is cumulative (e.g., #2 subsumes #1, #3 subsumes both #2 and #1, etc.). Thus, higher numbers are less restrictive.

1. Unlimited usage license (e.g., commercial use) – no restrictions on usage or distribution
2. Non-commercial license – modified source code may be redistributed – permits the distribution of modified source code and executables derived from modified source
3. Non-commercial license – includes source code, which may be modified for internal use
4. Non-commercial (research-only) license – for non-commercial academic use; free or reasonably-priced

When addressing the use cases listed above, please indicate licensing considerations for the following:

- Execution support technologies (e.g., operating systems, database systems, knowledge base systems, runtime libraries, etc.)
- Development support technologies (e.g., editors, knowledge base management tools, other proprietary tools used to develop program technologies, etc.)

A criterion for proposal evaluation will be the intellectual property licensing model. In general, less restrictive licenses are considered favorable. Since the research value of a deliverable depends upon the collective availability of its constituents, licensing models will be evaluated primarily in terms of their most restrictive provisions in the context of program goals.

As indicated previously, a key objective in this program is to promote the development and extension of machine reading technologies. Therefore, licensing models that avoid barriers to this objective will be viewed favorably. If the proposed licensing model poses restrictions that conflict with program goals, the offeror must recommend an amelioration strategy (e.g., open source alternatives, research versions, etc.).

Example

The following matrix depicts an example licensing model that is consistent with program goals. Each cell indicates one or more licenses that represent the intended level of restriction. The Open Source Initiative provides a formal definition of *open source* on their website at:

http://www.opensource.org/docs/definition.php

Licenses that fit this definition are represented in the matrix below as open source licenses or "OSL". Examples of conforming licenses are listed here:

http://www.opensource.org/licenses/

In addition, the Java Research License is available online at http://java.net/jrl.html. Note: OS = Operating System, and RTL = Run-time Library.

	Training & Evaluation Datasets	Overall System	Component Algorithms
Program Technology	OSL	OSL	OSL
Execution Support	OS and RTL: OEM license All other: JRL	OS and RTL: OEM license All other: JRL	OS and RTL: OEM license All other: JRL
Development Support	JRL	JRL	JRL

1.19 Human Use {No page limit}
For all proposed research that will involve human subjects in the first year or phase of the project, the institution must provide evidence of or a plan for review by an Institutional Review Board (IRB) upon final proposal submission to DARPA. For further information on this subject, see Section VI.B.4 below. If human use is not a factor in a proposal, then the offeror should state "NONE."

It is anticipated that some level of human use will be involved for the Evaluation Team offerors, but we do not expect any human use research for the Reading Team offerors.

1.20 Statement of Unique Capability Provided by Government Team Member {No page limit}
Per Section III.A. above, provide a statement which clearly demonstrates the work being provided by the Government entity team member is not otherwise available from the private sector. If none of the team members belongs to a Government entity, then the offeror should state "Not Applicable."

1.21 Government Team Member Eligibility {No page limit}
Per Section III.A. above, provide documentation, as required, citing the following specific information establishing a Government entity team member's eligibility to

propose to government BAAs: 1) statutory authority; 2) contractual authority; 3) supporting regulatory guidance; AND 4) evidence of agency approval. If none of the team members belongs to a Government entity, then the offeror should state "Not Applicable."

Proposal Section 2 - Additional Information {2 pages}
Offerors may submit a bibliography tailored specifically to show previous work relevant to this BAA. Note: This section is optional and will be considered for the reviewer's convenience only (i.e., will not be considered as part of the proposal for evaluation purposes).

Volume 2 – Cost Proposal

2. 1 Cover Sheet

- BAA number;
- Reading Team or Evaluation Team;
- Lead Organization Submitting proposal;
- Type of business, selected among the following categories: "LARGE BUSINESS", "SMALL DISADVANTAGED BUSINESS", "OTHER SMALL BUSINESS", "HBCU", "MI", "OTHER EDUCATIONAL", OR "OTHER NONPROFIT";
- Contractor's reference number (if any);
- Other team members (if applicable) and type of business for each;
- Proposal title;
- Technical point of contact to include: salutation, last name, first name, street address, city, state, zip code, telephone, fax (if available), electronic mail (if available);
- Administrative point of contact to include: salutation, last name, first name, street address, city, state, zip code, telephone, fax (if available), and electronic mail (if available);
- Award instrument requested: cost-plus-fixed-free (CPFF), cost-contract—no fee, cost sharing contract – no fee, or other type of procurement contract (*specify*), grant or cooperative agreement
- Place(s) and period(s) of performance;
- Total proposed cost separated by basic award and option(s) (if any);
- Name, address, and telephone number of the offeror's cognizant Defense Contract Management Agency (DCMA) administration office (*if known)*;
- Name, address, and telephone number of the offeror's cognizant Defense Contract Audit Agency (DCAA) audit office (*if known*);
- Date proposal was prepared;
- DUNS number;
- TIN number; and
- Cage Code;
- Subcontractor Information; and
- Proposal validity period.

2.2 Detailed cost breakdown {No page limit}
Provide: (1) total program cost broken down by major cost items (direct labor, including labor categories; subcontracts; costs to the government associated with any intellectual property licensing requirements; materials; other direct costs, overhead charges, etc.) and further broken down by task and phase; (2) major program tasks by fiscal year; (3) an itemization of major subcontracts and equipment purchases; (4) an itemization of any information technology (IT) purchase[1]; (5) a summary of projected funding requirements by month; and (6) the source, nature, and amount of any industry cost-sharing; and (7) identification of pricing assumptions of which may require incorporation into the resulting award instrument (e.g., use of Government Furnished Property/Facilities/Information, access to Government Subject Matter Expert/s, etc.). NOTE: for IT and equipment purchases, include a letter stating why the offeror cannot provide the requested resources from its own funding.

Provide supporting cost and pricing information in sufficient detail to substantiate the summary cost estimates, above. Include a description of the method used to estimate costs and supporting documentation. Note: "cost or pricing data" (as defined in FAR Subpart 15.4) shall be required if the offeror is seeking a procurement contract award of $650,000 or greater unless the offeror request an exception from the requirement to submit certified cost or pricing data. "Cost or pricing data" are not required if the offeror proposes an award instrument other than a procurement contract (e.g., a grant or cooperative agreement.)

The prime contractor is responsible for compiling and providing all subcontractor proposals for the Procuring Contracting Officer (PCO). Subcontractor proposals should include Interdivisional Work Transfer Agreements (ITWA) or similar arrangements. All proprietary subcontractor cost proposal documentation (prepared at the same level of detail as that required of the prime) which cannot be included with the prime's information, shall be made immediately available to the Government, upon request, under separate cover (i.e., mail, electronic/email, etc.), either by the offeror or by the subcontractor organization. See Section VI.B.8 for additional offeror responsibilities involving subcontracted efforts.

[1] IT is defined as "any equipment, or interconnected system(s) or subsystem(s) of equipment that is used in the automatic acquisition, storage, manipulation, management, movement, control, display, switching, interchange, transmission, or reception of data or information by the agency. (a) For purposes of this definition, equipment is used by an agency if the equipment is used by the agency directly or is used by a contractor under a contract with the agency which – (1) Requires the use of such equipment; or (2) Requires the use, to a significant extent, or such equipment in the performance of a service or the furnishing of a product. (b) The term "information technology" includes computers, ancillary, software, firmware and similar procedures, services (including support services), and related resources. (c) The term "information technology" does not include – (1) Any equipment that is acquired by a contractor incidental to a contract; or (2) Any equipment that contains imbedded information technology that is used as an integral part of the product, but the principal function of which is not the acquisition, storage, manipulation, management, movement, control, display, switching, interchange, transmission, or reception of data or information. For example, HVAC (heating, ventilation, and air conditioning) equipment such as thermostats or temperature control devices, and medical equipment where information technology is integral to its operation, are not information technology."

IV.C. Submission Dates and Times

The full proposal must be submitted per the instructions in Section IV.A above by 1200 noon (ET) on 30 December 2008 (initial closing), in order to be considered during the initial evaluation phase. While DARPA-BAA-09-03 will remain open until 1200 noon (ET) 13 November 2009 (final closing date/BAA expiration), offerors are warned that the likelihood of funding is greatly reduced for proposals submitted after the initial closing date.

DARPA will acknowledge receipt of complete submissions via email and assign control numbers that should be used in all further correspondence regarding proposals.

Failure to comply with the submission procedures may result in the submission not being evaluated.

IV.D. Intergovernmental Review - N/A

IV.E. Funding Restrictions

While DARPA currently anticipates using 6.2 funding for this program, if DARPA chooses to use 6.1 funding for any effort funded under this BAA, the Contractor is hereby notified that total negotiated indirect cost rates may not exceed 35% of the total cost of the award. Total costs include all bottom line costs. For Grant/Agreement awardees subject to the cost principles in 2 CFR part 220 (Educational Institutions), indirect costs are all costs of a prime award that are Facilities and Administration costs. For Grant/Agreement awardees subject to the cost principles in 2 CFR part 225 (State, Local, and Indian Tribal Governments), 2 CFR part 230 (Non-Profit Organizations) or 48 CFR part 32 (Federal Acquisition Regulation), indirect cost means any cost not directly identified with a single final cost objective, but identified with two or more final cost objectives or with at least one intermediate cost objective. The cost limitations do not flow down to subcontractors.

IV.F. Other Submission Requirements

Proposals MUST NOT be submitted to DARPA in hard copy (see Submission instructions above in Section IV.B).

University (prime) grant submissions may be made via the Grants.gov web site (http://www.grants.gov/) by using the "Apply for Grants" function. Duplicate submissions should not be uploaded to DARPA via the online tool described above in Section IV.B. However, offerors must still submit an online coversheet as described there.

V. APPLICATION REVIEW INFORMATION

V.A. Evaluation Criteria

Evaluation of proposals will be accomplished through a scientific review of each proposal using the following criteria. While these criteria are listed in descending order

of relative importance, it should be noted that the combination of all non-cost evaluation factors is significantly more important than cost.

The highest priority of evaluation criteria deals with the offerors' understanding of and justification to meet Go/No-Go requirements for each phase of the program. These metric criteria are specified in detail in the section entitled Evaluation Metrics and Assessment Protocols and Figure 8. This is essential in order to move from one phase into the next and will be determined at the end of each phase at a "Go/No-Go" session within DARPA. Offerors must clearly demonstrate their understanding of this fact within the proposal. At the completion of Phase 1, offerors must clearly pass the metric requirement outlined for this first phase which is significantly different from the metrics for the remaining phases. Since the metrics shift between Phase 1 and the rest of the Phases, at the end of Phase 1 offerors must demonstrate and justify their ability to complete the Phase 2 and subsequent phase metric requirements.

1. Ability to Meet Program Go/No-Go Metrics

The feasibility and likelihood of the proposed approach for satisfying the program Go/No-Go metrics are explicitly described and clearly substantiated. The proposal reflects a mature and quantitative understanding of the program Go/No-Go metrics, the statistical confidence with which they may be measured, and their relationship to the concept of operations that will result from successful performance in the program.

2. Overall Scientific and Technical Merit

The proposed technical approach is feasible, achievable, complete and supported by a proposed technical team that has the expertise and experience to accomplish the proposed tasks. Task descriptions and associated technical elements provided are complete and in a logical sequence with all proposed deliverables clearly defined such that a final product that achieves the goal can be expected as a result of award. The proposal identifies major technical risks and planned mitigation efforts are clearly defined and feasible.

3. Innovative Technical Solution to the Problem

The objective of this criterion is to establish that innovative and promising approaches are being applied to achieve the objectives of the effort. Offerors should apply new and/or existing theory and practice in an innovative way that supports the objectives of the proposed effort. The proposed approach concepts should show breadth of innovation across all the dimensions of the proposed solution. The theoretical enablers should be traceable to the objectives defined in the proposal.

4. Potential Contribution and Relevance to the DARPA Mission

The potential contributions of the proposed effort with relevance to the national technology base will be evaluated. Specifically, DARPA's mission is to maintain the technological superiority of the U.S. military and prevent technological surprise from harming our national security by sponsoring revolutionary, high-payoff research that

bridges the gap between fundamental discoveries and their military use. Since a key goal of the program is the advancement of science to maintain technological superiority, proposals should maximize the availability of machine reading technologies and data for reuse to further work in this field. This criterion will assess the extent to which the intellectual property stipulations of the proposal are consistent with the scientific advancement goals of the program and DARPA.

5. Offeror's Capabilities and/or Related Experience

The offeror's prior experience in similar efforts must clearly demonstrate an ability to deliver products that meet the proposed technical performance within the proposed budget and schedule. Similar efforts completed/ongoing by the offeror in this area are fully described including identification of other Government sponsors.

6. Plans and Capability to Accomplish Technology Transition

The capability to transition the technology to the research, industrial, and operational military communities in such a way as to enhance U.S. defense, and the extent to which intellectual property rights limitations creates a barrier to technology transition.

7. Realism of Proposed Schedule

The offeror's abilities to aggressively pursue performance metrics in the shortest timeframe and to accurately account for that timeframe will be evaluated, as well as offeror's ability to understand, identify, and mitigate any potential risk in schedule.

8. Cost Realism

The objective of this criterion is to establish that the proposed costs are realistic for the technical and management approach offered, as well as to determine the offeror's practical understanding of the effort. This will be principally measured by cost per labor-hour and number of labor-hours proposed. The evaluation criteria recognize that undue emphasis on cost may motivate offerors to offer low-risk ideas with minimum uncertainty and to staff the effort with junior personnel in order to be in a more competitive posture. DARPA discourages such cost strategies. Cost reduction approaches that will be received favorably include innovative management concepts that maximize direct funding for technology and limit diversion of funds into overhead.

NOTE: OFFERORS ARE CAUTIONED THAT EVALUATION RATINGS MAY BE LOWERED AND/OR PROPOSALS REJECTED IF SUBMITTAL INSTRUCTIONS ARE NOT FOLLOWED.

V.B. Review and Selection Process

It is the policy of DARPA to ensure impartial, equitable, comprehensive proposal evaluations and to select the source (or sources) whose offer meets the Government's technical, policy, and programmatic goals. Pursuant to FAR 35.016, the primary basis for selecting proposals for acceptance shall be technical, importance to agency programs, and fund availability. In order to provide the desired evaluation, qualified

Government personnel will conduct reviews and (if necessary) convene panels of experts in the appropriate areas.

Proposals will not be evaluated against each other since they are not submitted in accordance with a common work statement. DARPA's intent is to review proposals as soon as possible after they arrive; however, proposals may be reviewed periodically for administrative reasons. For evaluation purposes, a proposal is the document described above in IV.B.2. Other supporting or background materials submitted with the proposal will be considered for the reviewer's convenience only and not considered as part of the proposal.

Award(s) will be made to offerors whose proposals are determined to be the most advantageous to the Government, all factors considered, including the potential contributions of the proposed work to the overall research program and the availability of funding for the effort. Award(s) may be made to any offeror whose proposal is determined selectable regardless of its overall rating.

Restrictive notices notwithstanding, offerors are advised that employees of support contractors under contract to the Government may be used by DARPA to administratively process proposals or perform other administrative duties requiring access to other contractors' proprietary information. These support contracts include nondisclosure agreements prohibiting their contractor employees from disclosing any information submitted by other contractors or using such information for any purpose other than that for which it was furnished. By submission of its proposal, each offeror agrees that proposal information may be disclosed to those non-Government personnel for the limited purposes stated above. In addition, these support contractors are prohibited from competition in DARPA technical research. Subject to the restrictions set forth in FAR 37.203(d), input on technical aspects of the proposals may be solicited by DARPA from non-Government consultants/experts who are strictly bound by the appropriate non-disclosure requirements.

It is the policy of DARPA to treat all proposals as competitive information and to disclose their contents only for the purpose of evaluation. No proposals will be returned. Upon completion of the source selection process, the electronic version of each proposal submitted will be retained at DARPA.

VI. AWARD ADMINISTRATION INFORMATION

VI.A. Award Notices

As soon as the evaluation of a proposal is complete, the offeror will be notified that 1) the proposal has been selected for funding pending contract negotiations, or, 2) the proposal has not been selected. These official notifications will be sent via US mail to the Technical POC identified on the proposal coversheet.

VI.B.Administrative and National Policy Requirements

VI.B.1. Security Classification and Proprietary Issues

Security classification guidance on a DD Form 254 (DoD Contract Security Classification Specification) will not be provided at this time, since DARPA is soliciting ideas only and does not encourage classified proposals in response to this announcement. However, after reviewing incoming proposals, if a determination is made that contract award may result in access to classified information, a DD Form 254 will be issued upon contract award. ***If you choose to submit a classified proposal you must first receive the permission of the Original Classification Authority to use its information in replying to this announcement.*** NOTE: If proposals are classified, the proposals must indicate the classification level of not only the proposal itself, but also the anticipated award document classification level. Applicable classification guide(s) should also be submitted to ensure the proposal is protected at the appropriate classification level.

All proposals containing proprietary data should have the cover page and each page containing proprietary data clearly marked as containing proprietary data. It is the offeror's responsibility to clearly define to the Government what is considered proprietary data.

The Government anticipates proposals submitted under this BAA will be unclassified. However, if a proposal is submitted as "Classified National Security Information" (as defined by Executive Order 12958 as amended), then the information must be marked and protected as though classified at the appropriate classification level and then submitted to DARPA for a final classification determination.

Classified submissions shall be appropriately and conspicuously marked with the proposed classification level and declassification date. Submissions requiring DARPA to make a final classification determination shall be marked as follows:

> CLASSIFICATION DETERMINATION PENDING. Protect as though classified (insert the recommended classification level: (e.g., Top Secret, Secret or Confidential)

Classified submissions shall be in accordance with the following guidance:

Confidential and Secret Collateral Information: Use classification and marking guidance provided by previously issued security classification guides, the Information Security Regulation (DoD 5200.1-R), and the National Industrial Security Program Operating Manual (DoD 5220.22-M) when marking and transmitting information previously classified by another Original Classification Authority. Classified information at the Confidential and Secret level may be mailed via appropriate U.S. Postal Service methods (e.g., (USPS) Registered Mail or USPS Express Mail). All classified information will be enclosed in opaque inner and outer covers and double wrapped. The inner envelope shall be sealed and plainly marked with the assigned classification and addresses of both sender and addressee. The inner envelope shall be addressed to:

Defense Advanced Research Projects Agency
ATTN: Information Processing Techniques Office
Reference: DARPA-BAA-09-03
3701 North Fairfax Drive
Arlington, VA 22203-1714

The outer envelope shall be sealed with no identification as to the classification of its contents and addressed to:

Defense Advanced Research Projects Agency
Security & Intelligence Directorate, Attn: CDR
3701 North Fairfax Drive
Arlington, VA 22203-1714

All Top Secret materials: Top Secret information should be hand carried by an appropriately cleared and authorized courier to the DARPA CDR. Prior to traveling, the courier shall contact the DARPA CDR at 571 218-4842 to coordinate arrival and delivery.

Special Access Program (SAP) Information: SAP information must be transmitted via approved methods. Prior to transmitting SAP information, contact the DARPA SAPCO at 703-526-4052 for instructions.

Sensitive Compartmented Information (SCI): SCI must be transmitted via approved methods. Prior to transmitting SCI, contact the DARPA Special Security Office (SSO) at 703-248-7213 for instructions.

Offerors must have existing and in-place prior to execution of an award, approved capabilities (personnel and facilities) to perform research and development at the classification level they propose. It is the policy of DARPA to treat all proposals as competitive information, and to disclose their contents only for the purpose of evaluation. Proposals will not be returned. The original of each proposal received will be retained at DARPA and all other non-required copies destroyed. A certification of destruction may be requested, provided the formal request is received at this office within 5 days after unsuccessful notification.

VI.B.2. Intellectual Property

VI.B.2.a. Procurement Contract Offerors

VI.B.2.a.i Noncommercial Items (Technical Data and Computer Software)

Offerors responding to this BAA requesting a procurement contract to be issued under the FAR/DFARS shall identify all noncommercial technical data and noncommercial computer software (including open source and open source modifications) that it plans to generate, develop, and/or deliver under any proposed award instrument in which the Government will acquire less than

unlimited rights, and to assert specific restrictions on those deliverables. Offerors shall follow the format under DFARS 252.227-7017 for this stated purpose. In the event that offerors do not submit the list, the Government will receive "unlimited rights" to all noncommercial technical data and noncommercial computer software generated, developed, and/or delivered under any award instrument, unless it is substantiated that development of the noncommercial technical data and noncommercial computer software occurred with mixed funding. If mixed funding is anticipated in the development of noncommercial technical data and noncommercial computer software generated, developed, and/or delivered under any award instrument, then offerors should identify the data and software in question, as subject to Government Purpose Rights (GPR). In accordance with DFARS 252.227-7013 Rights in Technical Data - Noncommercial Items, and DFARS 252.227-7014 Rights in Noncommercial Computer Software and Noncommercial Computer Software Documentation, the Government will automatically assume that any such GPR restriction is limited to a period of five (5) years in accordance with the applicable DFARS clauses, at which time the Government will acquire "unlimited rights" unless the parties agree otherwise. Offerors are admonished that the Government will use the list during the source selection evaluation process to evaluate the impact of any identified restrictions and may request additional information from the offeror, as may be necessary, to evaluate the offeror's assertions. If no restrictions are intended, then the offeror should state "NONE."

A sample list for complying with this request is as follows:

NONCOMMERCIAL			
Technical Data Computer Software To be Furnished With Restrictions	Basis for Assertion	Asserted Rights Category	Name of Person Asserting Restrictions
(LIST)	(LIST)	(LIST)	(LIST)

VI.B.2.a.ii Commercial Items (Technical Data and Computer Software)

Offerors responding to this BAA requesting a procurement contract to be issued under the FAR/DFARS shall identify all commercial technical data and commercial computer software (including open source) that may be embedded in any noncommercial deliverables contemplated under the research effort, along with any applicable restrictions on the Government's use of such commercial technical data and/or commercial computer software. In the event that offerors do not submit the list, the Government will assume that there are no restrictions on the Government's use of such commercial items. The Government may use the list during the source selection evaluation process to evaluate the impact of any identified restrictions and may request additional information from the offeror, as may be necessary, to evaluate the offeror's assertions. If no restrictions are intended, then the offeror should state "NONE."

A sample list for complying with this request is as follows:

COMMERCIAL			
Technical Data Computer Software To be Furnished With Restrictions	Basis for Assertion	Asserted Rights Category	Name of Person Asserting Restrictions
(LIST)	(LIST)	(LIST)	(LIST)

VI.B.2.b. Non-Procurement Contract Offerors – Noncommercial and Commercial Items (Technical Data and Computer Software)

Offerors responding to this BAA requesting a grant or cooperative agreement shall follow the applicable rules and regulations governing these various award instruments, but in all cases should appropriately identify any potential restrictions on the Government's use of any Intellectual Property contemplated under those award instruments in question. This includes both Noncommercial Items and Commercial Items. Although not required, offerors may use a format similar to that described above. The Government may use the list during the source selection evaluation process to evaluate the impact of any identified restrictions, and may request additional information from the offeror, as may be necessary, to evaluate the offeror's assertions. If no restrictions are intended, then the offeror should state "NONE."

VI.B.2.c. All Offerors – Patents

Include documentation proving your ownership of or possession of appropriate licensing rights to all patented inventions (or inventions for which a patent application has been filed) that will be utilized under your proposal for the DARPA program. If a patent application has been filed for an invention that your proposal utilizes, but the application has not yet been made publicly available and contains proprietary information, you may provide only the patent number, inventor name(s), assignee names (if any), filing date, filing date of any related provisional application, and a summary of the patent title, together with either: 1) a representation that you own the invention, or 2) proof of possession of appropriate licensing rights in the invention.

VI.B.2.d. All Offerors – Intellectual Property Representations

Provide a good faith representation that you either own or possess appropriate licensing rights to all other intellectual property that will be utilized under your proposal for the DARPA program. Additionally, offerors shall provide a short summary for each item asserted with less than unlimited rights that describes the nature of the restriction and the intended use of the intellectual property in the conduct of the proposed research.

VI.B.3. Meeting and Travel Requirements

There will be a program kickoff meeting and all key participants are required to attend. Performers should also anticipate periodic site visits at the Program Manager's discretion.

VI.B.4. Human Use

All research involving human subjects, to include use of human biological specimens and human data, selected for funding must comply with the federal regulations for human subject protection. Further, research involving human subjects that is conducted or supported by the DoD must comply with 32 CFR 219, *Protection of Human Subjects* (http://www.dtic.mil/biosys/downloads/32cfr219.pdf), and DoD Directive 3216.02, *Protection of Human Subjects and Adherence to Ethical Standards in DoD-Supported Research* (http://www.dtic.mil/whs/directives/corres/html2/d32162x.htm).

Institutions awarded funding for research involving human subjects must provide documentation of a current Assurance of Compliance with Federal regulations for human subject protection, for example a Department of Health and Human Services, Office of Human Research Protection Federal Wide Assurance (http://www.hhs.gov/ohrp). All institutions engaged in human subject research, to include subcontractors, must also have a valid Assurance. In addition, personnel involved in human subjects research must provide documentation of completing appropriate training for the protection of human subjects.

For all proposed research that will involve human subjects in the first year or phase of the project, the institution must provide evidence of or a plan for review by an Institutional Review Board (IRB) upon final proposal submission to DARPA. The IRB conducting the review must be the IRB identified on the institution's Assurance. The protocol, separate from the proposal, must include a detailed description of the research plan, study population, risks and benefits of study participation, recruitment and consent process, data collection, and data analysis. Consult the designated IRB for guidance on writing the protocol. The informed consent document must comply with federal regulations (32 CFR 219.116). A valid Assurance, along with evidence of appropriate training for all investigators, should accompany the protocol for review by the IRB.

In addition to a local IRB approval, a headquarters-level human subjects regulatory review and approval is required for all research conducted or supported by the DoD. The Army, Navy, or Air Force office responsible for managing the award can provide guidance and information about their component's headquarters-level review process. Note that confirmation of a current Assurance and appropriate human subjects protection training is required before headquarters-level approval can be issued.

The amount of time required to complete the IRB review/approval process may vary depending on the complexity of the research and/or the level of risk to study participants. Ample time should be allotted to complete the approval process. The IRB approval process can last for one to three months, followed by a DoD review that can last for three to six months. No DoD/DARPA funding can be used toward human subjects research until ALL approvals are granted.

VI.B.5. Animal Use

Any Recipient performing research, experimentation, or testing involving the use of animals shall comply with the rules on animal acquisition, transport, care, handling, and use in: (i) 9 CFR parts 1-4, Department of Agriculture rules that implement the Laboratory Animal Welfare Act of 1966, as amended, (7 U.S.C. 2131-2159); (ii) the guidelines described in National Institutes of Health Publication No. 86-23, "Guide for the Care and Use of Laboratory Animals"; (iii) DoD Directive 3216.01, “Use of Laboratory Animals in DoD Program.”

For submissions containing animal use, proposals should briefly describe plans for Institutional Animal Care and Use Committee (IACUC) review and approval. Animal studies in the program will be expected to comply with the PHS Policy on Humane Care and Use of Laboratory Animals, available at http://grants.nih.gov/grants/olaw/olaw.htm.

All Recipients must receive approval by a DoD certified veterinarian, in addition to an IACUC approval. No animal studies may be conducted using DoD/DARPA funding until the USAMRMC Animal Care and Use Review Office (ACURO) or other appropriate DoD veterinary office(s) grant approval. As a part of this secondary review process, the Recipient will be required to complete and submit an ACURO Animal Use Appendix, which may be found at https://mrmc.amedd.army.mil/AnimalAppendix.asp

VI.B.6. Publication Approval

It is the policy of the Department of Defense for products of fundamental research to remain unrestricted to the maximum extent possible. Contracted fundamental research is defined as research performed under grants and contracts that are (a) Basic Research, whether performed by universities or industry or (b) applied research and performed on-campus at a university. The research shall not be considered fundamental in those rare and exception circumstances where the applied research effort presents a high likelihood of disclosing performance characteristics of military systems or manufacturing technologies that are unique and critical to defense, and where agreement on restrictions have been recorded in the contract or grant.

It is anticipated that the performance of research resulting from this BAA is expected to be fundamental research.
Offerors are advised if they propose grants or cooperative agreements, DARPA may elect to award other award instruments. DARPA will make this election if it determines that the research resulting from the proposed program will present a high likelihood of disclosing performance characteristics of military systems or manufacturing technologies that are unique and critical to defense. Any award resulting from such a determination will include a requirement for DARPA permission before publishing any information or results on the program. The following provision will be incorporated into any resultant non-fundamental research procurement contract:

> There shall be no dissemination or publication, except within and between the Contractor and any subcontractors, of information developed under this contract or contained in the reports to be furnished pursuant to this contract without prior

written approval of the DARPA Technical Information Officer (DARPA/TIO). All technical reports will be given proper review by appropriate authority to determine which Distribution Statement is to be applied prior to the initial distribution of these reports by the Contractor. Papers resulting from unclassified contracted fundamental research are exempt from prepublication controls and this review requirement, pursuant to DoD Instruction 5230.27 dated October 6, 1987.

When submitting material for written approval for open publication, the Contractor/Awardee must submit a request for public release to the DARPA TIO and include the following information: 1) Document Information: document title, document author, short plain-language description of technology discussed in the material (approx. 30 words), number of pages (or minutes of video) and document type (briefing, report, abstract, article, or paper); 2) Event Information: event type (conference, principle investigator meeting, article or paper), event date, desired date for DARPA's approval; 3) DARPA Sponsor: DARPA Program Manager, DARPA office, and contract number; and 4) Contractor/Awardee's Information: POC name, e-mail and phone. Allow four weeks for processing; due dates under four weeks require a justification. Unusual electronic file formats may require additional processing time. Requests can be sent either via e-mail to tio@darpa.mil or via 3701 North Fairfax Drive, Arlington VA 22203-1714, telephone (571) 218-4235. Refer to www.darpa.mil/tio for information about DARPA's public release process.

VI.B.7. Export Control

Should this project develop beyond fundamental research (basic and applied research ordinarily published and shared broadly within the scientific community) with military or dual-use applications the following apply:

- The Contractor shall comply with all U. S. export control laws and regulations, including the International Traffic in Arms Regulations (ITAR), 22 CFR Parts 120 through 130, and the Export Administration Regulations (EAR), 15 CFR Parts 730 through 799, in the performance of the contract or agreement. In the absence of available license exemptions/exceptions, the Contractor shall be responsible for obtaining the appropriate licenses or other approvals, if required, for exports (including deemed exports) of hardware, technical data, and software, or for the provision of technical assistance.
- The Contractor shall be responsible for obtaining export licenses, if required, before utilizing foreign persons in the performance of this contract, including instances where the work is to be performed on-site at any Government installation (whether in or outside the United States), where the foreign person will have access to export-controlled technologies, including data or software.
- The Contractor shall be responsible for all regulatory record keeping requirements associated with the use of licenses and license exemptions/exceptions.
- The Contractor shall be responsible for ensuring that the provisions of this clause apply to its subcontractors.

VI.B.8. Subcontracting

Pursuant to Section 8(d) of the Small Business Act (15 U.S.C. 637(d)), it is the policy of the Government to enable small business and small disadvantaged business concerns to be considered fairly as subcontractors to contractors performing work or rendering services as prime contractors or subcontractors under Government contracts, and to assure that prime contractors and subcontractors carry out this policy. Each offeror who submits a contract proposal and includes subcontractors is required to submit a subcontracting plan in accordance with FAR 19.702(a) (1) and (2) should do so with their proposal. The plan format is outlined in FAR 19.704.

VI.B.9. Central Contractor Registration (CCR)

Offerors selected, but not already registered in the Central Contractor Registry (CCR) will be required to register in CCR prior to any award under this BAA. Information on CCR registration is available at http://www.ccr.gov

VI.B.10. On-line Representations and Certifications (ORCA)

In accordance with FAR 4.1201, prospective offerors shall complete electronic annual representations and certifications at http://orca.bpn.gov.

VI.B.11. Wide Area Work Flow (WAWF)

Unless using another approved electronic invoicing system, performers will be required to submit invoices for payment directly via the Internet/WAWF at http://wawf.eb.mil. Registration to WAWF will be required prior to any award under this BAA.

VI.C. Reporting

VI.C.1. T-FIMS

The award document for each proposal selected and funded will contain a mandatory requirement for four DARPA/IPTO Quarterly Status Reports each year, one of which will be an annual project summary (a final report that summarizes the project and tasks, notwithstanding the fact that the research may be continued under a follow-on vehicle). These reports may be electronically submitted by each awardee under this BAA via the DARPA Technical – Financial Information Management System (T-FIMS). The T-FIMS URL and instructions will be furnished by the contracting agent upon award. In addition, each performing contractor (including subs) on each team will be expected to provide monthly status reports to the Program Manager. Reports and briefing material will also be required as appropriate to document progress in accomplishing program metrics. There may also be additional reporting requirements for grants and cooperative agreements.

VI.C.2. I-Edison

All required reporting shall be accomplished, as applicable, using the i-Edison.gov reporting website at http://s-edison.info.nih.gov/iEdison

VII. AGENCY CONTACTS

DARPA will use electronic mail for all technical and administrative correspondence regarding this BAA, with the exception of selected/not-selected notifications. Administrative, technical or contractual questions should be sent via e-mail to DARPA-BAA-09-03@darpa.mil. If e-mail is not available, please fax questions to (703) 741-0228, Attention: Machine Reading Solicitation. All requests must include the name, email address, and phone number of a point of contact.

Solicitation Web site: http://www.darpa.mil/ipto/solicit/solicit.asp.

VIII. OTHER INFORMATION

The solicitation web page at http://www.darpa.mil/ipto/solicit/solicit.asp will have additional information and a Frequently Asked Questions (FAQ) list.

VIII.A. Industry Day

An industry day is planned in conjunction with this BAA. It is scheduled for November 21, 2008 in Arlington, VA. Additional details may be found at http://www.csc-ballston.com/darpa/registration/intro.asp?regCode=CgWCaAgU. Interested parties must be registered on this site no later than November 20, 2008.

VIII.B. References

VIII.B.1. Terminology Appendix

This section provides definitions and/or greater detail on concepts and artifacts discussed elsewhere in this solicitation. This information varies in depth of coverage and is not intended to be definitive or prescriptive. Rather, it is intended to help provide insight into DARPA's thinking at the time this BAA was written. The descriptions of procedures, interfaces and other artifacts are expected to evolve and change in response to inputs and requirements from the selected performer teams.

Knowledge Forms and Representations

The goal of MRP is to develop general-purpose technology for mapping from a wide range of natural texts onto formal reasoning systems. A wide range of reasoning mechanisms, with a corresponding range of representations, has been successfully used to gain high performance in narrow domains. The MRP aims to provide knowledge for these domains *expressed in the reasoning systems most appropriate for those domains.* In addition, the MRP hopes to foster a number of competing reading technologies, which implies multiple representations may be used by the Reading Systems as well. Therefore the program doesn't wish to constrain the internal representational choices of either the Reading System or the DSRS.

Yet the Reading System and the performance context of the DSRS must communicate for two purposes: First, the Reading System must be able to convey to the DSRS the relevant knowledge it has accumulated by reading, so that a performance task may be

accomplished. Second, the Reading System must be able to take advantage of the task-specific knowledge possessed by the DSRS to assist in the reading process itself, without requiring each Reading System to incorporate such knowledge into its own structures.

The solution adopted by the MRP is to use a single, relatively straightforward language for representing target knowledge. *Target knowledge* is all of the knowledge that the Reading System reads and infers from the text corpus, including alternative interpretational hypotheses. This target knowledge may be passed between the Reading System and the DSRS in order to help the Reading System disambiguate what it has read. Target knowledge may also be passed between the DSRS and the Reading System so that knowledge absorbed by the Reading System can be used to help the DSRS perform correctly.

A Reading System may represent, index, and manage knowledge (including text, assertions, and inferential closures) in any way desired. However, this knowledge must ground out as (or be translatable to) statements in the target knowledge language.

The language defined below utilizes familiar formal-reasoning concepts. It is based on the well-defined framework of RDF from the Semantic Web, which allows for the partial specificity of statements, a desirable attribute. A number of concepts are defined below. Note that the representation language and interface described here are initial versions; they are expected to evolve throughout the MRP.

Statements

The lowest level of the target knowledge language is the statement.

> A *statement* is simply an RDF triple. An *RDF triple* has three arguments in the format (subject predicate object) and is used to represent logical statements such as “the predicate holds for the subject” (with null object) or “the property (or attribute) called predicate of subject has a value of object”.

The *predicate* is a relational term in the RDF language. In the first phases of the program, all of these relational terms will be captured in the Syntax Specification and Mapping (discussed below) and made available to the Reading System. In later phases the Reading System will be expected to learn and create new relational terms.

The *subject* and *object* are Skolem, or global, terms. The Reading System is expected to extend this set of entities and classes. The syntax or semantics of a subject, predicate, or object are not constrained further, in order to enable the broadest interactions using target knowledge.

Assertions

The next level of the target knowledge language is the assertion.

> An *assertion* is a set of statements (also known as an *RDF graph*), plus an associated correctness probability. The *correctness probability* represents the certainty that all of the statements are true. This is a syntactic, not

semantic, specification chosen to enable interaction with many different inference mechanisms and representations.

As an example, if the Reading System read the sentence, “Pietro lives in Arlington, Virginia.”, it might produce this corresponding assertion:

90%: { <residesAt Pietro PietroHouse>, <inCity PietroHouse Arlington>, <inState Arlington Virginia> }

Note that the correctness probability is an internal assessment, by either the Reading System or DSRS, of how likely the conjunction of statements is true. This assessment can be derived in any way each system desires.

In Phase 1 of the MRP, there will be no representation of time or modality. These will be added to the language in later phases.

Formal Knowledge

Formal knowledge in the target knowledge representation is made up of assertion sets that encode a set of assertions that the Reading System deems to be sufficiently consistent and relevant that it is appropriate to ask the DSRS to perform inference over them. As the Reading System processes more text, it will typically grow a number of assertion sets that are internally consistent, but not necessarily consistent with (or related to) each other. It is the Reading System’s task to manage the creation, deletion, merger, and splitting of assertion sets as processing progresses.

Inferential Knowledge

The formal knowledge consists of ground facts and inferential knowledge. *Inferential knowledge* consists of rules of inference.

Acquired Inferential Knowledge – *Acquired inferential knowledge* consists of rules of inference that are acquired by reading, either from an explicit textual representation of the rule or by generalization from many examples of the rule’s use in texts. Each such rule affects the performance of a Reading System by inferring new assertions from the assertions extracted directly from text.

Acquired inferential knowledge (or acquired rules) could contribute to the overall inferential behavior of a Reading System and affect its behavior over very large assertion sets. In later phases of the program, acquiring such rules by reading will be addressed.

The Reading System will not be responsible for capturing all inferential knowledge implied by available texts, rather only specific, limited classes of such knowledge that commonly occur. Possible examples of such rules:

- task-specific rules that fill in information “immediately implied” by read facts;
- data-consistency rules (e.g., one cannot be in two places at the same time, but can have two friends at the same time);

- frame axioms, which determine how the truth values of all assertions are updated as a function of a change in the world;
- strategy rules for deciding when and how to apply read assertions, to avoid inferring all possible things.

Teams will propose the forms of inferential knowledge they will acquire from text and the approach(es) they will consider. A representation for inferential knowledge is not presented here as this is expected to vary across the Reading Teams.

Inputs and Interfaces

Figure 9 is a somewhat simplified version of Figure 2 that focuses in the inputs to the Reading Systems and the interfaces between the Reading Tasks and Reading Systems.

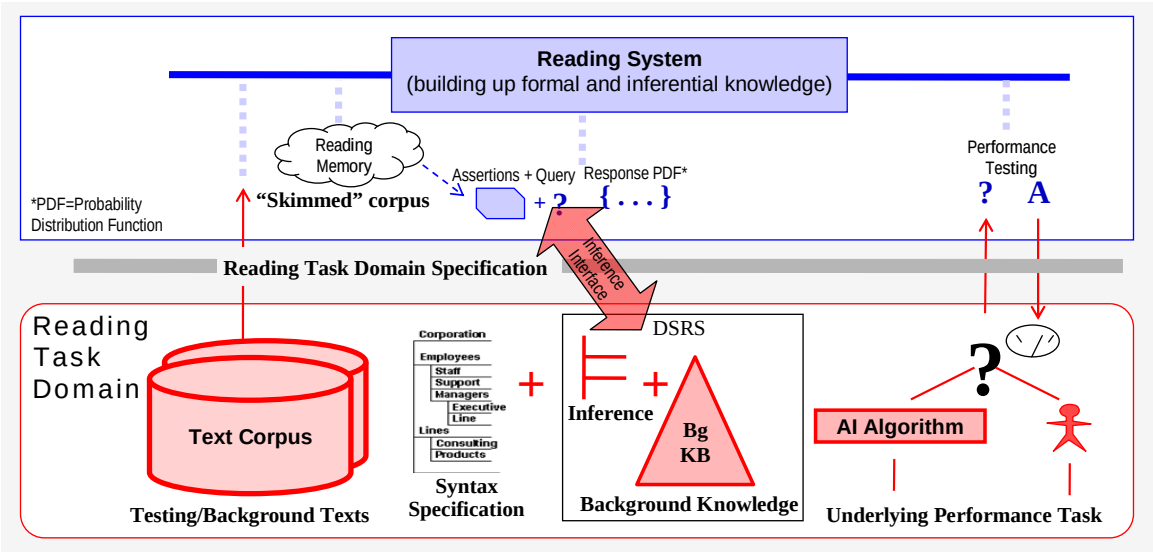

Figure 9 Communications between Reading Tasks and Reading Systems

Reading Task Domain

The set of components that collectively define a reading and reasoning context is referred to as a Reading Task Domain. This includes the testing and background corpora, one or more Syntax Specifications and Mappings, a DSRS for accessing task-specific background knowledge, and a set of performance-task queries with consensus results. The one-to-many mapping between testing corpus and Syntax Specifications and Mappings enables a single corpus to support multiple performance tasks, using focused subsets of larger theories, with independently measureable testing targets. For example, a single corpus about organizations might support one reasoning task with queries about management chains as well as a different reasoning task about workers' physical locations. The components within each reading task domain will conform to the reading task domain specification.

Text Corpus

Corpora will be formatted as UTF-8 documents with very limited markup. Markup elements may be limited to titles and section headers but is not expected to include full XHTML or even Dublin Core metadata. Metadata provided will represent some "least common denominator" compatible with many text sources. As with other artifacts described in this section, the specific implementation details are expected to evolve over the course of the MRP.

Testing Corpus

Reading Systems are responsible for assertions and inferences that can be drawn from information contained in the testing corpus. This is the primary source for knowledge obtained by reading (as distinct from knowledge available inherently through the DSRS.) These are natural English texts from a variety of sources that vary in style, complexity, etc. as appropriate for the task domains and program phase.

Background Corpus (Optional)

This optional input is designed to widen the spectrum of approaches that may be attempted within the MRP. The background corpus is an optional corpus that has some of the same characteristics as the testing corpus (drawn from the same or a similar source and discussing the same or similar topics) but is ***not*** required for answering the evaluation questions. Rather this optional corpus may be provided to facilitate approaches such as those that use large amounts of unlabeled text or statistically-based learning approaches in the training of Reading Systems, without forcing non-statistical, "deep parsing" approaches to wade through large corpora (particularly early in the program) during the evaluation task. Reading Teams may use or ignore this corpus as they deem appropriate. While these texts will have similar style, complexity, etc. as the testing corpora, Reading Systems will not require any information contained in these optional texts to respond to the evaluation queries. Thus, teams will *not* be responsible for information contained in these texts.

Domain-Specific Reasoning System (DSRS)

Notionally, a DSRS is a "black box" that is capable of performing inferences relevant to the associated reading task domain. This is also the vehicle by which reading systems access task-specific background knowledge. The DSRS is queried using the inference interface.

Task-Specific Background Knowledge – Task-specific knowledge that the Evaluation Team and DARPA deem necessary for the reading and reasoning task and that is made available in the DSRS. In particular, this is the knowledge that would either not be found in natural texts, or is knowledge that would be out of scope for extraction by reading. Task-

specific background knowledge, regardless of the form it takes, is expected to be supplied by the DSRS and made available via the inference interface.

Syntax Specification and Mapping – The *Syntax Specification* specifies the syntactic structure of the formal knowledge used in the target AI task. The *Mapping* provides a set of natural text sentences and phrases with the same meaning as the formal assertions.

The Syntax Specification by itself is essentially a query ontology that defines the syntax for target tasks. It is expressed in the form of a RDF Schema, or RDFS [see http://www.w3.org/TR/rdf-schema/].

In addition to specifying the syntactic representations used in the formal target task, this specification also specifies example mappings from formal knowledge to English sentences and phrases with an equivalent meaning. These example mappings provide correspondences between the representation of the formal problem requests, as supplied by the problem source, and the natural text corpora used to answer those queries. For each term in the ontology, an explicit mapping is provided from that term to a small number of example sentences or phrases.

As an example, this mapping might assert that: "Defense Advanced Research Projects Agency" maps onto the formal RDFS category Organization. Or it might assert that the tagged English text: "[Dr. Tether $_{person}$] is now the longest serving [DARPA $_{organization}$] director." asserts: <WorksFor DrTether$_{person}$ DARPA$_{organization}$>.

In Phase 1 of the MRP, the Syntax Specification and Mapping will completely specify the knowledge representations that must be read and encoded in order to solve the reasoning tasks. In later phases this is a "starter" representation that must be extended through specialization or augmented with other ontologies. In the final program phases completely new representations must be derived from the texts.

Performance-Task Queries and Consensus Results

Recall that the goal of machine reading is to enable a performance task to be executed. Requests or queries for a performance task will use the same query template employed by the inference interface, and the results will be a set of probabilities over ground instances just as with the inference interface.

In the case of a performance-task request, however, no assertion set is provided; only a query template is given. Instead, the entire test corpus is the informational basis for answering the performance-task requests. A major challenge for the Reading System will be to decide *which* parts of the texts and which text-to-knowledge translations should be used to respond to each request. Reading Systems will need to employ measures of internal inferential consistency, contextual clues of relevancy, and other such techniques. Teams should propose techniques for managing this mismatch between formal knowledge and natural text corpora.

Since Go/No-Go metrics are based on comparisons to human performance on the same task, a consensus set of human response assertions to the queries will also be produced for use in assessing the correctness and completeness of the answers produced by the Reading Teams for those queries. This source may be a manually constructed set of questions and associated set of answers or it may be some AI reasoning system whose execution requires queries to be handled by the Reading System based on the textual information.

Reading Task Domain Specification

The reading task domain specification is a formal interface that allows the specification of all context information required to provide a new problem domain to a Reading System. This interface provides access to:

- The testing and background text corpora;
- A Syntax Specification and Mapping of the performance task queries over which the system is to be tested;
- A DSRS for this domain, containing both initial background knowledge and a reasoning capability appropriate to the domain;
- An inference interface to the DSRS;
- Performance task queries and associated consensus responses.

The first three of these are discussed above as components of the Reading Task Domain itself. The latter two are discussed below.

Inference Interface – The *inference interface* is the software interface between the Reading System and the DSRS that allows the Reading System to successfully read domain-specific text and the DSRS to perform task-specific inference that relies upon knowledge gleaned from text. The result set returned by the DSRS may also specify an associated weighting or probability expressing the confidence in each resulting assertion. In addition to providing probabilistic results, an entailment operator (see *probabilistic entailment operator* discussion below) will also supply a derivation tree rooted in the provided assertions and background knowledge. The goal of decoupling a general purpose reading capability from domain-specific reasoning, using this inference mechanism as the sole bridge between the two, may not be achievable in practice. Approaches that require deeper integration with the inference mechanism will be considered, but the tradeoff between the stand-alone generality of the reading technology and its dependence on the reasoning context will be weighed. As the program progresses and linguistic constructs for modal and temporal reasoning are considered, this entailment operator will be generalized to support that complexity.

The primary mechanism by which the inference interface works is a probabilistic entailment operator. Before discussing this some additional terminology must be introduced.

Query Template – A *query template* is an assertion containing one or more statements (the *query statements*) that have one or more unbound variables (called *template variables*) in lieu of the usual RDF-triple arguments. A query

template is used to specify one or more statements for which truth values are sought, as well as one or more statements that are assumed to be true with the accompanying correctness probability and used to establish the context for making inferences to answer the query.

In many domains, a great many assertions may follow from an initial set of assertions. Thus, it will be much more effective and computationally efficient to have the Reading System specify in a query template which arguments of which statements should be derived (if possible) and the set of context-setting statements that should be assumed in so doing. A query template specifies a possibly infinite set of assertions that are of interest.

Syntactically, a query template is encoded as an assertion where one or more arguments in the query statement are replaced by template variables (denoted by a variable name preceded by a question mark). An assertion, ***A***, is said to *match* a query assertion, ***Q***, if ***A*** can be derived from ***Q*** by a consistent substitution of RDF terms for all of the template variables in ***Q***.

A query template may be used by the Reading System to query the DSRS.

Query – A *query* is the submission of a query template by the Reading System to the DSRS via the inference interface.

Result Set – A *result set* is the assertion set that is returned as the response to a query. All of the assertions in the result set match the query template and have the stated correctness probability. Thus, a result set captures the limited set of assertions that are implied by the query statements.

As an example, a task-specific inference rule in the DSRS might say that a person usually works for the same company as his or her manager. A Reading System might read the sentence “Bob is at DARPA and manages John.” and hypothesize that it implies:

80%: { <manages Bob John>, <employedBy Bob DARPA> }.

The DSRS could then be given that single assertion as a trivial assertion set and the following query template:

{ <employedBy John ?x> }.

The result set would probably contain a single assertion:

{ 80%: { <employedBy John DARPA> } **}**.

As a richer example, imagine reading about personal travel itineraries. One might read that Jim is giving a talk at Northwestern University at 10:30 a.m. this Wednesday, and then read a sentence that appears to say that he will be attending a lecture at George Mason University at 1:00 p.m. on Wednesday. The assertions that the Reading System directly derives from the two sentences appear to it to be

compatible. Only the application of reasoning about the travel domain will show that they are actually incompatible. First, the fact that Jim is presenting at Northwestern implies that he will be in Chicago, Illinois, on Wednesday. Second, he must also be in Virginia around the same time on Wednesday, since George Mason is in Fairfax, Virginia. Finally, the minimum travel time between these two cities precludes both statements being true. A rich task-specific inference mechanism could provide a general interface for incorporating such specialized knowledge into a general Reading System.

Returning now to the definition of the Probabilistic Entailment Operator and its role in the inference interface:

> **Probabilistic Entailment Operator** – The *probabilistic entailment operator*, given a query template, returns a result set. The result set lists the assertions that follow from the query template and the available background knowledge in the DSRS and that also match the query statements. The result set also specifies a correctness probability.
>
> In addition to providing probabilistic results, the entailment operator also supplies a derivation tree for the results, rooted in the given assertions and the background knowledge utilized. As the program progresses and linguistic constructs for modal and temporal reasoning are considered, this entailment operator will be generalized to match the added complexity.

The goal of decoupling a general-purpose reading capability from domain-specific reasoning, using this inference mechanism as the sole bridge between the two, may not be achievable in practice. Reading System developers may need to extend or change the background knowledge. Approaches that require deeper integration with the inference mechanism (providing deeper insight into the inferences made) will be considered, but the value gained by having reading technology that is dependent on the reasoning context will be weighed against the loss of generality implied by affording such access. It is anticipated that over the course of this program one or more “sweet spots” in this tradeoff curve of generality versus power will be identified.

VIII.B.2. Example Tasks Appendix

In this section we provide an example of a readability assessment task to be provided in Phase 1, and several reading tasks which show the progressive complexity that will be explored in this program. These are intended to be illustrative, not prescriptive. The Evaluation Team will develop (with concurrence from DARPA) the actual tasks used in the MRP.

VIII.B.2.a. Readability Assessment Example

It is assumed that a suitable benchmark for readability is based upon subjective “expert” human assessment. An informal study conducted at DARPA suggests that strong agreement among educated humans exists about the readability of a passage of text.

Thus, machine-based readability assessment success is defined in terms of making an assessment that is closer to the average human assessment than the mean human error.

The following three passages on the topic of “Knowledge Management” have been assessed by humans for readability:

TEXT A

> *Unfortunately, there's no universal definition of knowledge management (KM), just as there's no agreement as to what constitutes knowledge in the first place. For this reason, it's best to think of KM in the broadest context. Succinctly put, KM is the process through which organizations generate value from their intellectual and knowledge-based assets. Most often, generating value from such assets involves codifying what employees, partners and customers know, and sharing that information among employees, departments and even with other companies in an effort to devise best practices. It's important to note that the definition says nothing about technology; while KM is often facilitated by IT, technology by itself is not KM.*

TEXT B

> *In an organizational context, data represents facts or values of results, and relations between data and other relations have the capacity to represent information. Patterns of relations of data and information and other patterns have the capacity to represent knowledge. For the representation to be of any utility it must be understood, and when understood the representation is information or knowledge to the one that understands. The value of Knowledge Management relates directly to the effectiveness with which the managed knowledge enables the members of the organization to deal with today's situations and effectively envision and create their future. Without on-demand access to managed knowledge, every situation is addressed based on what the individual or group brings to the situation with them. With on-demand access to managed knowledge, every situation is addressed with the sum total of everything anyone in the organization has ever learned about a situation of a similar nature.*

TEXT C

> *There is a broad range of thought on Knowledge Management with no unanimous definition. The approaches vary by author and school. Knowledge Management may be viewed from each of the following perspectives:*

- *Techno-centric: A focus on technology, ideally those that enhance knowledge sharing and growth.*
- *Organizational: How does the organization need to be designed to facilitate knowledge processes? Which organizations work best with what processes?*
- *Ecological: Seeing the interaction of people, identity, knowledge and environmental factors as a complex adaptive system.*

Knowledge Management has always existed in one form or another. Examples include on-the-job peer discussions, formal apprenticeship, discussion forums, corporate libraries, professional training and mentoring programs. However, with computers becoming more widespread in the second half of the 20th century, specific adaptations of technology such as knowledge bases, expert systems, and knowledge repositories have been introduced to further enhance the process.

The following table shows example ratings of the three textual passages on a Likert scale of 1 to 3, where 1 = “low readability”, 2=”moderate readability”, 3=”high readability”. For each passage, the mean human rating and mean error for human ratings is given.

	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Mean	Mean Error
TEXT A	2	2	2	2	1	1.8	0.32
TEXT B	1	1	1	1	3	1.4	0.64
TEXT C	3	3	3	3	2	2.8	0.32
Mean error across samples:							**0.427**

The next two tables depict the ratings of two fictitious Reading Systems, one which demonstrates success in assessing readability, and other that fails.

Reading System that passes

	MR rating	Mean Human Rating	Error (diff)
TEXT A	2	1.8	0.2
TEXT B	2	1.4	0.6
TEXT C	3	2.8	0.2
			0.33 (< 0.427)

Reading System that fails

	MR rating	Mean Human Rating	Error (diff)
TEXT A	3	1.8	1.2
TEXT B	2	1.4	0.6
TEXT C	1	2.8	1.8
			1.2 (> 0.427)

VIII.B.2.b. Examples of Data Capture from Text

Following the Phase 1 readability assessment challenge, the MRP will explore reading tasks of increasing complexity. The first tier in this exploration involves several variants of data capture from text. In particular, the MRP seeks to achieve sophisticated data capture (or “data extraction”, “knowledge capture”, “knowledge extraction”) that requires bringing multiple types of knowledge to bear: local textual information, global textual context, multiple sources of data (e.g., text, formatted data), and knowledge of the domain of discourse. Another MRP goal is for Reading Systems to handle multiple types of naturally occurring text (e.g., textbooks, email messages, dialogs). To highlight the features of this problem space, this section describes four successively more difficult types of data capture that a Reading System should be able to accomplish over the course of the program:

- The first example is of the capture of simple facts, each contained within a single sentence;
- The second example discusses extraction from heterogeneous types of data, such as a message or report that consists of both relational information (e.g., message headers, document metadata) and textual content;
- The third example presented is of understanding broad summary concepts from a large stream of data;
- The fourth example is presented as a stretch problem in which a diagnostic system learns heuristics from reading transcripts of earlier diagnosis sessions.

VIII.B.2.b.i Capture of Localized Relational Data

Beyond data capture from text, a higher echelon of complexity involves the capture of localized relational data.

Consider a simple example involving a passage of source text about employment relationships.

> Bob is an office director at DARPA. One of Bob’s employees, John, has been working on a BAA. The BAA will be reviewed by Dave, one of John’s colleagues, before it is released. Dave will certainly be talking to Jim about the role his employer, Department of Commerce, will play.

The Reading System is provided access to some inferential knowledge (via a DSRS) that can answer questions probabilistically about this domain, as described earlier in the BAA. In processing this passage (and ignoring the temporal aspects for now), the Reading System should be able to determine:

> <hasRole Bob Manager>, <employedBy John Bob>,
> <coworker John Dave>, <isA employer DepartmentOfCommerce>

With these assertions and background knowledge about employment relationships, the system should be able to determine that John, Dave, and Bob are all persons and that

DARPA is an employer. Furthermore, the Reading System could now query the DSRS for plausible assertions about Dave based upon the assertions derived from the passage. This query would return the assertion that, with high probability, <employedBy Dave DARPA> is true.

This new inference provided by the DSRS now permits the Reading System to resolve the pronoun ambiguity in the sentence "Dave will certainly be talking to Jim about the role his employer, Department of Commerce, will play." Since <employedBy Dave DARPA> is probably true, then it is most likely that "his" refers to "Jim", and therefore <employedBy Jim DepartmentOfCommerce> is true.

Thus, mapping linguistic relationships derived from the original text into relations that are precisely defined within the DSRS on employment resulted in a new assertion that enabled the Reading System to further refine its parsing. This combination of linguistic and domain inference has been used to capture more data from the corpus than would have been possible without the benefit of the domain knowledge.

Since specific instances of the relations in both systems have some degree of uncertainty, the Reading System may generate other plausible parses and ask the DSRS if they make sense. For each such parse, the DSRS may return a probabilistic distribution of alternative sets of consistent relations. At this point, the Reading System will likely have a number of hypotheses, each consisting of a set of statements that are consistent from both the linguistic and domain viewpoints. However, the example set discussed above has a probability much higher than any other set, so all the others could be pruned away at an appropriate point.

The DSRS may also have constraint rules that apply to its objects, attributes, and relations. If the sentence, "John is a colleague of Todd." were read, the DSRS would return <coworker John Todd> as consistent with <coworker John Dave> because an employee is allowed to have many coworkers. However, if the sentence, "John works for Alice." were now read, the DSRS might return a number of possibilities. First, it would create a hypothesis that includes statements that John reports to Alice and no longer reports to Bob. It would also suggest that John might be an indirect report to Alice (i.e., Alice is Bob's boss) and is still a direct report to Bob. Of course, this discussion omits the consideration of time, since John might have reported to Bob earlier and now reports to Alice. Handling time will be required in later phases of the MRP.

This simple example has demonstrated how domain-specific inference can be integrated into and used in domain-independent reading in a general way.

VIII.B.2.b.ii Capture of Data from Heterogeneous Sources

Another high-complexity reading task is the capture of data from heterogeneous sources. The central benefit of structured data is the ability to formally process that source. So assessing trends, averages, generalizations, outliers, and many other patterns are greatly aided when data is captured formally. Practical repositories

however, often associate unstructured textual information (e.g., a variable-length text field in a RDBMS) along with the formally expressed knowledge as a “catch all” for important information that is not expressible in the designed formal structure. Allowing the advantages of the formal processing to extend to the textual knowledge contained in these catch-all text fields would be of great value. Thus, another target for the MRP would be the exploitation of heterogeneous knowledge sources, that is, knowledge sources that combine text and relational data about the same topic.

For example, military after-action reports (AARs) typically have both highly formatted fields and free-text fields. The highly formatted fields may include reporting unit, type of unit, time, location, author, a code for type of incident encountered (e.g., sniper fire, improvised explosive device [IED], mortar fire, etc.), enemy unit, type of enemy unit, size of enemy unit, type of IED used, number of fatalities on both sides, number of injured, etc. The free-text fields describe in arbitrary detail more information about the reporting unit, the location of the event, the surroundings and situation, type of IED, the enemy unit, specific individuals involved, etc. Although some value can be derived by performing statistical analysis on the formatted data, much of the information most useful to commanders and intelligence analysts is found in the text fields.

The AARs described above are typically stored in either a custom database or a commercial relational database. In either case the formatted fields can be searched with an SQL query (or something similar), while the text fields can be searched by keywords and phrases (and sometimes Boolean combinations). A single search yields a subset of records, which must be examined manually to determine their actual relevance.

One can imagine an “IED reporter”, a special-purpose inference system that knows how to evaluate and create a fairly detailed analysis report on IED events. The purpose of a Reading System would be to walk through the AAR database and feed relevant information about IED events to the IED reporter. Since the IED reporter would have no way of telling the Reading System what to read and what to ignore, the front-end would have to read every field of every record. It could use the mechanism described in the last section for processing text-only documents to process the textual entries in the database. However, it would also have to ingest formatted data as well, which would require inferring the name of the object (typically the first field of the record) and the name of the attribute (the name of the column). It would then have to understand the text entries in the context of the rest of the record. For example, there might be a column called “Event Details”, whose entry is a text description of the type of event, time, and location given in those fields of the same record.

Another factor making heterogeneous data a difficult problem is the fact that fields can have a range of precise semantics, from highly precise and relatively obvious (e.g., a non-negative integer representing the number of events that occurred) to the most imprecise (e.g., arbitrary natural language). In the former case, the semantics can be embedded in the DSRS. Of greater difficulty is the case where a field lies somewhere in the middle of the semantic-precision scale. For example, a name field may allow an

arbitrary set of characters, and so could contain just the last name, only the first name, a compound last name, or a full name including middle name(s) or initial(s). Although this particular example could also occur in a text-only document, other such “limited text” fields may have highly idiosyncratic vocabulary, syntax, and semantics, which will require a special-purpose parser to handle properly and a learning system to configure properly. General purpose technologies that can be easily adapted to handling these special cases robustly would be needed to handle this broad class of reading task domains.

VIII.B.2.b.iii Capture of Data from a Data Stream

In addition to capturing localized relational data and heterogeneous data, another significant class of data sources where MR technologies would have great impact involves capturing data from text streams. These streams could take the form of newswire text, a progression of email messages, blogs, wiki updates, etc., or a series of intelligence reports over time. There are three attributes of such stream data that must be handled by the Reading System.

First, the “granularity” of the reading task expands greatly with streamed data. In the section on “Capture of Localized Relational Data”, a new piece of information was always contained within a single sentence and examined in the context of a small number of facts gleaned from a few preceding sentences. In a single continuous stream however, there may be large amounts of text (e.g., from thousands of news articles) from different sources and by different authors. Therefore, the same concept needs to be understood, even though it occurs in two documents that are thousands of “pages” and many hours apart and written by different people using different vocabularies and writing styles. Additionally, facts derived from a variety of places in the stream may provide the convergent knowledge necessary to constrain the set of interpretations and finally understand some concept.

As an example of how domain knowledge can be used to bridge the understanding gaps that can occur between related, but isolated (in time and position) sentences, consider the following sentence (translated from the source in the German media) and the corresponding extracted knowledge:

> “In the Saarland, the **governing CDU** was able to remain in power, the SPD lost seven seats, while the Liberals and Greens were able to re-enter state parliament.”

Upon analysis by the Reading System, this text might give rise to the following triples (among others):

> (is_a CDU organization)
> (has_role CDU governing)
> (location CDU Saarland)

A separate story from the German media, arriving much later, contains the following sentence:

> "At the federal level, the **Christian Democrats** lost their dominance but remain **part of the governing coalition**.”

This text might result in these triples:

(is_a Christian_Democrats organization)
(has_role Christian_Democrats governing)
(location Christian_Democrats federal_level)

Though a relationship exists between these two sentences, there is nothing in the text to reveal it. One possible way to determine that “CDU” and “Christian Democrats” are both short forms for the same formal name is to notice that both are governing organizations, Saarland is a state in Germany, “federal level” refers to Germany (from elsewhere in the second article), and that both proper nouns include a capital “C” and a capital “D”. However, this analysis is somewhat strained.

A better approach might be to figure out that both articles are discussing German politics and then search for more information in other sources. For example, Wikipedia provides the following additional information:

> "Politics of Germany takes place in a framework of a federal parliamentary representative democratic republic, whereby the Federal Chancellor is the head of government, and of a plurality multi-party system. Since 1949, the party system has been dominated by the **Christian Democrats (CDU)** and the Social Democrats (SPD).”

This text would lead to the following triples, which clearly indicate the logical equivalence of the nouns "CDU" and "Christian Democrats":

(is_a Christian_Democrats organization)
(same_as CDU Christian_Democrats)

A third way to determine this equivalence would be to ask a DSRS with political expertise what else is known about “CDU” and “Christian Democrats”. If it knows about politics in Germany, it would probably know that both “CDU” and “Christian Democrats” are shorthand for the Christian Democratic Union, the party’s official name.

A second problem created by data streams is the introduction of time. Over a time period of almost any duration, important states of the world that were initially true may no longer be true, and vice versa. To deal with streamed data, the Reading System will have to incorporate some sort of temporal reasoning capability so that it will know at what point in time or during what interval in time a statement held true.

A third problem created by data streams is their perpetually evolving nature. This implies that the potential scope of analysis is unlimited. Not only may individual assertions change multiple times over a longer period of time, the relevant chunk sizes of time may change. For example, one indicator of a country’s instability is its economic, social, and political trends. The US is vitally interested in what these trends are and how fast they are changing. However, an Army patrol in Baghdad may be most concerned about what has changed in one neighborhood in the last 24 hours, while the Defense Intelligence Agency in Washington, DC, is more interested in how Iraq, as a whole, has changed over the past year. A Reading System, therefore, needs to possess an ability to adapt and handle a variety of time spans in its analyses.

VIII.B.2.c. Natural Language–Based Diagnosis Engine

The previous section provided examples of how machine reading could be applied to isolated chunks of text, multimedia (text and relational) data, and streamed text to populate a knowledge base with basic facts, summaries, and trends. This section introduces a much more difficult reading problem: that of creating a natural language–based diagnosis engine that gains its expertise by reading and learning diagnostic heuristics from transcripts of a human expert who has conducted real diagnostic sessions at a help desk. This type of knowledge is clearly much more specialized and more of a problem-solving nature. This task represents the most difficult sort of challenge the MRP might tackle and may be considered a “stretch” goal of the program. A problem like this would only be considered in the final phases of the program.

Knowledge-based diagnostic systems have long been an important application of AI technology, going back to expert systems such as MYCIN. Typically, the challenge is to infer (primarily by abduction) the cause of some abnormal behavior, given background knowledge about a system or device, a description of the faulty system’s outputs, and the corresponding input stimuli. Background knowledge starts with a model of the system being diagnosed, its component elements, its interconnections, and its behavior. Background knowledge may also include a model of the abnormal behaviors and their causes (a fault model) and a diagnosis strategy.

Diagnosis is an exemplar of a class of non-trivial inference systems that could be integrated with a general-purpose Reading System. So, although the inferences required are specific to diagnosis, the types of rules to be extracted from the text are still reasonably general purpose in kinds and nature. The use of diagnosis is intended to push to the limit what can reasonably be expected of a general-purpose Reading System.

For the diagnosis problem, the Reading System will again be supplied with three external inputs. There will be a DSRS containing a sophisticated ontology about the class of devices to be diagnosed, and a Syntax Specification and Mapping to the DSRS. The background knowledge in the DSRS will be incomplete and unable to perform diagnosis without further information. This information will come from a set (possibly in the hundreds or thousands) of natural-language trouble reports and/or transcripts of diagnostic sessions. Trouble reports contain a description of a problem, the set of diagnostic steps taken, and the outcome. Transcripts are the verbatim dialog between a device user with a problem and a diagnostic expert who is located at a remote call center. A transcript contains descriptions of abnormal behavior, requests by the diagnostician for troubleshooting tests to be performed by the user, results of the tests relayed by the user back to the diagnostician, diagnoses of the problem, repair actions, and additional tests to confirm that the repair was successful.

These reports and transcripts will allow the Reading System to learn the variety of abnormal behaviors and their possible causes. The Reading System must also employ the background knowledge of the DSRS to help derive these rules. The resulting

performance engine will be capable of performing diagnoses. It will be tested by posing novel diagnosis problems to it and then assessing the correctness of the resulting diagnoses.

The Reading Teams will again be given a description of the probabilistic entailment operator to be used to query the provided diagnosis DSRS. For this task the operator will have some very specialized inferential capabilities. In particular, it will accept a sequence of diagnostic actions involving, say, part x and part y, and return the probability that part x functionally depends on part y, or even what relationship might exist between x and y. For example, the notion that x will only function if y is also functioning might be captured by the relation <functionallyRequired x y>. These inferences are very specific to one domain (e.g., PC diagnosis) and therefore are not expected to be handled by the Reading System. However, given appropriate background knowledge, such a relation might be inferable from, say, the order that questions were asked by during one the transcripts that were read. Assuming such generic inferential knowledge is captured about diagnosis, then a generic Reading System can make such diagnosis specific inferences even though reading technology developed in not specialized to the diagnosis task. Note further that with such inferential knowledge the surface form of the text, and the knowledge ultimately derived from that text, can be quite removed from one another, given such deep background knowledge.