Andrew. J. Cowell, Deborah L. McGuinness, Carrie F. Varley, and David A. Thurman. Knowledge-Worker Requirements for Next Generation Query Answering and Explanation Systems. In the Proceedings of the Workshop on Intelligent User Interfaces for Intelligence Analysis, International Conference on Intelligent User Interfaces (IUI 2006), Sydney, Australia.

Knowledge-Worker Requirements for Next Generation Query Answering & Explanation Systems

Andrew J. Cowell[a], Deborah L. McGuinness[b], Carrie F. Varley[c], David A. Thurman[c]

[a]Comp. & Info. Sci. Directorate
Battelle-Pacific NW Division
andrew@pnl.gov

[b]Knowledge Systems Laboratory
Stanford University
dlm@ksl.stanford.edu

[c]National Security Directorate
Battelle-Pacific NW Division
{carrie.varley,dave}@pnl.gov

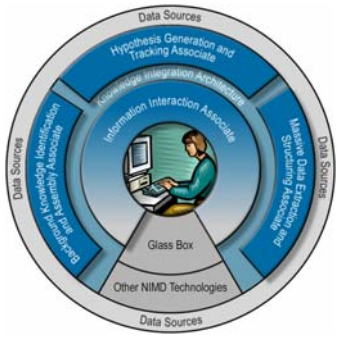
ABSTRACT
Knowledge workers need tools to help them navigate through, evaluate, and understand large stores of information. Motivated by the needs of ARDA's Novel Intelligence from Massive Data program, Battelle, Stanford University, and IBM have developed a suite of technologies for knowledge discovery, knowledge extraction, knowledge representation, automated reasoning, explanation, and human-information interaction. Our team has developed an integrated analytic environment composed of a collection of analyst associates, software components that aid the analyst at different stages of the analytical process, collectively known as "Knowledge Associates for Novel Intelligence (KANI)." As part of this effort, we have incorporated a Query Answering and Explanation component that allows analysts to pose questions of the system based on the knowledge it has of a particular domain and specific tasking (problem). Answers are presented along with optional information about sources, assumptions, explanation summaries, and interactive justifications. This paper describes the analyst requirements and response to the explanation component of the KANI system. We believe the explanation infrastructure, its interface for analysts and knowledge workers, and the provenance requirements are all contributions that can be leveraged beyond the KANI implementation.

Categories and Subject Descriptors
H.5.2 [**User Interfaces**]: *Evaluation/methodology; Interaction styles; Theory and methods; User-centered design.*

General Terms
Design, Experimentation, Human Factors,

Keywords
Query Answering & Explanation Systems; User-Centered Design; Focus Groups; Requirements Elicitation; Reasoning Systems.

1. INTRODUCTION
The Advanced Research and Development Activity (ARDA) sponsors high-risk, high-payoff research into advanced information technologies addressing issues faced by the intelligence community. ARDA's Novel Intelligence from Massive Data (NIMD) program aims to assist analysts in successfully coping with the volumes and varieties of data inundating them. While NIMD focuses on intelligence community analysts, the tasks performed by such analysts are similar to those performed by knowledge workers in other areas, such as business analysts and scientific researchers who work with large, complicated, inter-related data sets and make decisions informed by combinations of knowledge from disparate sources. As such, many of the approaches and technologies developed under NIMD will likely have broad applicability. The Knowledge Associates for Novel Intelligence (KANI) project team[1] is developing a system of automated "associates" to actively support and participate in the intelligence analysis task. Their role is to help analysts identify, structure, aggregate, analyze, and visualize task-relevant information and to help them construct explicit models of alternative hypotheses (scenarios, relationships, causality, etc.). The KANI associates also actively assist an analyst in analytical reasoning such as hypothesis refinement, contradiction detection, and assumption testing. The primary enabler of these capabilities is the production and use of computer interpretable and explainable knowledge expressed in formal knowledge representation languages and the design of knowledge integration technologies that make KANI a natural part of the analyst's work processes. Figure 1 shows a conceptual architecture of the analytic support environment consisting of four knowledge associates and an encompassing knowledge integration architecture:



[1] The team is comprised of members from Battelle Pacific Northwest Division, IBM's T.J. Watson Research Center, and Stanford University Knowledge Systems, Artificial Intelligence Laboratory.

Andrew. J. Cowell, Deborah L. McGuinness, Carrie F. Varley, and David A. Thurman. Knowledge-Worker Requirements for Next Generation Query Answering and Explanation Systems. In the Proceedings of the Workshop on Intelligent User Interfaces for Intelligence Analysis, International Conference on Intelligent User Interfaces (IUI 2006), Sydney, Australia.

Figure 1: The KANI Conceptual Architecture

The Hypothesis Generation and Tracking Associate assists analysts by guiding and accelerating the analytic processes that are orchestrated and directed by the analyst. The Massive Data Extraction and Structuring Associate ingests text documents, identifies and extracts relevant knowledge, provides structured annotations and ontologies, and provides that information to other KANI associates. The Background Knowledge Identification and Assembly Associate enables analyst to identify and assemble relevant structure, semi-structured, and unstructured background information for a given set of documents and a given task through semantic search techniques tailored to models of prototypical analytic tasks. Finally, the Information Integration Associate (IIA) facilitates analyst interaction with each of the other KANI associates and provides interactive representations of the analytic process that can be inspected, revised, shared, explained, and analyzed for patterns, biases, and deficiencies.

2. QUERY ANSWERING & EXPLANATION SYSTEMS

Explanation systems have a long history and have possibly been made most famous in the expert systems implementations. Any system that provides answers to user questions may eventually face questions as to why a user should believe an answer. In the early days of explanation systems, a typical scenario included expert input of data and integrated, trustworthy systems. Still, complicated systems had significant explanation requirements in terms of helping users to understand how conclusions were reached. The explanations in those systems typically focused on some (understandable) presentation of a reasoning trace. One early prototypical example system was the MYCIN [1] system that diagnosed infectious diseases and could explain its reasoning. This led to work on Teiresias that was built to help refine the MYCIN knowledge and thus further expose the reasoning and the EMYCIN work that helped provide a foundation for generating expert systems and included some explanation capabilities. Another generation of explanation systems was introduced with the Explainable Expert System [2] when systems were designed with explanation in mind. These systems however all typically had the same assumptions – that data was reliable, rules (once deployed and debugged) were reliable, and question answering systems and reasoners were integrated and reliable. The distributed and evolving nature and the diversity of the web has broken all of these assumptions. In worlds such as the one KANI exists in, not all data sources are reliable or current, reasoning techniques (such as extractors) are not all sound and complete and question answering systems may be quite distributed and composed of components with varying degrees of testing and reliability. Thus, today's explanation systems require a much broader range of support in terms of including information about sources, methods, dates, etc., in addition to the traditional summaries of execution traces. Our explanation solution embodied in the KANI effort is based on the Inference Web [3]. It attempts to address the diversity of today's explanation needs.

3. INTERVIEW

In the rest of the paper, we report on interviews with analysts that were aimed at capturing their explanation needs for question answering systems such as KANI. We will describe our knowledge capture process and our findings.

3.1 Method

Here we discuss the method used to conduct the interviews. The setting was informal (a standard conference room) and lasted three hours, with some email and in-person follow-up.

3.1.1 Lexicon Capture

One of the most challenging aspects of dealing with a user community is 'speaking their language.' The session began with 15 minutes of conversation around the topic of QA&E to ascertain the lexicon used by the analysts. For example, we wanted to understand how they talk about evidence, queries, answers, explanations, evidentially, provenance, the nomenclature used and how formally they referred to each.

3.1.2 Context of Use Analysis

In order to understand the tools and techniques that analysts use to pose questions to systems—both computerized (e.g., Goggle) and human (e.g., a collections department or librarian)—as well as how they deal with the answers retrieved and explanations (if any) offered, we undertook a 'context of use' (CoU) analysis. CoU is a structured method for eliciting detailed information about a product, procedure or methodology and how it is used to perform a familiar task [4]. The approach required the construction and posing of certain probe questions about how the analysts currently use QA&E technologies in order to bring this information out into the open for discussion and to focus thoughts on QA&E prior to exploring a KANI-specific scenario. CoU analysis is often administered as a type of interactive questionnaire but in order to engage the analysts, a focus group setting was used.

3.1.3 Scenarios

Following the CoU analysis, a KANI usage scenario was presented. Scenarios (in general) are characterizations of users and their tasks in a specified context (e.g., in this case, the performance of a specific task within the KANI environment). They offer concrete representations of a user working with a product in order to achieve a particular goal. This session was presented in a lecture-type format, and while clarifying questions were allowed (in order to not lose anyone), detailed discussions about KANI functionality not directly related to QA&E was not entertained.

3.1.4 Brainstorming

A brainstorming session was introduced in context of the scenario previously presented. Brainstorming is used to generate new ideas by freeing the mind to accept any idea that is suggested, thus allowing freedom for creativity. Through discussion and sketches, the analysts were encouraged to envision how they might utilize a query and explanation capability within KANI. This included main concepts such as how they would

envision interacting with a next generation query answering and explanation system. They were encouraged to identify both functional concepts (e.g., "I would want to see…", "There should be a way to…") and presentation concepts (e.g., "I'd like to see a button here that would…").

3.1.5 Wizard of Oz

The final segment of the session was planned to consist of a Wizard of Oz session. Wizard of Oz is a technique used to present advanced concepts of interactions to users. In essence, the session organizer plays the role of 'computer' and 'processes' input from the user and emulates system output. The aim is to demonstrate computer capabilities and to clarify the results of the brainstorming session. Unfortunately, time ran out and the team was unable to perform this segment.

3.2 Participant Profiles

Table 1 describes the profiles of the three Battelle analysts that took part in the study.

Table 1. Profiles of Battelle Analysts Partaking in the Study

Analyst	A	B	C
Gender	F	M	F
Age	30s	40s	30s
Handedness (L or R)	R	Both	L
Education	M.S. Cog Psych.	M.S. Comp. Sci.	BSW (Social Work)
Years Experience	5	3	2
Specialty	Network Traffic	Network Traffic	Intelligence Analysis
Prior Occupation	UI Design & Human Factors	Computer Security	Social Work

3.3 Results

3.3.1 Asking the Query

The analysts talked of a number of systems that they currently use, from the ubiquitous Google web search engine to very specialized database search tools. Generally, their method of submitting queries was the same across the board (entry of text into a text box), with some requiring a specific syntax—either delimited responses for key terms like author (au=) or title (ti=) or query language (e.g., SQL92)—while others allowed a more free form approach (e.g., Google). Most analysts have specific training on how to best form queries (e.g., through the use of delimiters to ensure words are or are not within the solution set, Boolean operators, regular expressions, etc.) and these strings can often be fairly long. In the case of standing queries (i.e., search terms that the analyst uses regularly to be kept aware of new information) the analysts set up search profiles that they can request to be ran on a schedule (e.g., nightly, weekly or monthly). For example, a country analyst may need to keep up to date about a particular field in relation to their country of choice, and by setting up a search profile, any new hits will be delivered to them (usually through email). Google Alerts provide a similar functionality.

When interacting with a next generation intelligent QA&E system, the analysts expressed a preference to use similar mechanisms. This is not surprising as individuals generally gravitate to what feels natural to them, even when what they believe to be *natural* is due only to extended exposure. For example, all analysts had a desire to interact using traditional keyword search techniques instead of utilizing a (conceptually) more organic natural language system. Their reasoning was due to past experiences with such systems where the underlying understanding mechanism was unable to truly capture what the analysts were trying to ask, and hence provided sub par results. One analyst mentioned that she sometimes would use a natural language query if she what having difficulties determining a sufficiently powerful set of keywords. The documents returned would help in determining what keywords should then be used to obtain useful search results. Another issue with such next generation systems is being able to understand the corpus over which the system has knowledge. Many search engine sites also provide a directory view where individuals can browse instead of searching and the analysts envisioned that similar functionality would be an essential element in being able to understand what lies behind the advanced QA&E system.

Being able to expand, automatically, any and all keywords to take into consideration relevant synonyms and antonyms was seen as a timesaver – currently this is done manually if at all.

Customization of the environment was mentioned as being important. Being able to set up your environment so that you could move between machines (that may have access to different sources of information) and be able to interact in the same manner across all workstations was seen as very desirable.

3.3.2 Presenting the Results

The current tools used by the analysts present their results sets in a similar fashion. Just as Google presents a list of results, sorted by relevance, the other sources and databases provided a similar view of search results. Some enable the analyst to preview certain pertinent information such as the classification level of the material, the date it was published, the title of the piece, the author and what agency they belong to. Other more detailed metadata is also available, but was infrequently used by our analysts. Some systems (e.g., Google) allow immediate access to the original source documents and even to cached versions that provide a snapshot of how the information looked at a particular time in the past. Google also highlights occurrences of keywords within the document so that it is easy to determine the context.

When interacting with next generation QA&E systems, the analysts expect to see a more interactive, semantically rich environment. Disambiguation was a topic of great interest and analysts foresaw a step between the query being submitted and the solutions being presented where they would help direct the system in specifying the correct semantics. The system would therefore know to present results in the context of China the country, not a type of pottery. Eventually, as a system comes to understand the types of query an analyst is likely to ask, this step could be refined with a suggested context and eventually (potentially) avoided.

As keywords are the main element in determining what results are presented, analysts would enjoy being able to turn on and off certain keywords. This would enable them to review the 'solution dynamics' without having to resubmit the query numerous times. It would also allow them to get some idea of the importance of specific keywords and how the opinions expressed are directly related to the occurrence of the keyword.

Instead of just listing the documents from a solution set, the analysts talked of a hierarchical mechanism that would sort the returned material according to user-specified criteria. This could be multi-dimensional, presenting all documents of a particular type together (e.g., PDF files, Microsoft Word files, Microsoft Powerpoint files, etc.) or presenting results groupings (e.g., this set matched on keyword 1, 2 and 3). The analysts expect to see a mixture of modalities and be able to preview them inline (i.e., instead of a line of text describing a map, they should be able to see a thumbnail of a map).

The analysts expect to be able to sort the solution set via any of the available metadata – that is, by date, document title, relevance (usually equated to a score), author, agency, trust (as calculated by the system or as annotated in meta data), etc. In addition, they would like to annotate the results and save their query answering sessions in order to revisit past analyses. Of specific importance to the analysts was the ability to mark those documents that were used and those not used in a particular tasking. Another annotation dimension suggested was the ability to mark material as 'good' or 'bad' (which could in turn be used to train the system and help in presenting high impact results first). These annotations may be desirable to be maintained as private information available only to the analyst who entered the information.

3.3.3 Evaluating the Results/Explanations

The first evaluation of a document occurs prior to its content being physically accessed. It occurs at the metadata level. The credentials of who has captured/supplied/authored the material are first evaluated to provide the analyst with some idea of how reliable it may be. One of the analysts kept a specific record of the individuals she trusted, while the other relied on memory and subjective judgment. This extends to the organization or agency that individual may represent. Thus, information concerning the source of the information and the agencies either employing this source or providing some validation of the source are important.

Additionally, the analysts stated that citations of sources were an important part in judging the reliability and trustworthiness of the source. The date is often the next item to be analyzed. Assuming the tasking is not an historical study, the date may indicate that the material is outdated and/or potentially obsolete. Interestingly, the title was one of the least important features for these analysts. Of greater importance are the keywords the analyst used to generate the query and knowing that they are present in the document.

A large part of the analytical tradecraft involves comparing material. Analysts often look for differences between reports from different sources, looking for facts that may be mentioned in one document but not in another. In the rush to publish the story first, early articles may not be the most accurate. Later versions may not include specific information because it has been found to be wrong or considered irrelevant (it may also have been removed for more clandestine reasons).

When the results come from electronic sources, such as web pages or web logs (blogs), analysts actively look for elements of credibility within information available. They evaluate the URL and base their initial evaluation on where the information is coming from (for example, they usually assume a URL ending in .gov to be credible, although an example was discussed where a government site had been hacked. Additional rules of thumb were discussed as indicators of potential hacking such as spelling and grammar errors). They consider trust on an institutional, as opposed to individual, basis (especially if they have had no previous experience with that individual). Thus, they will typically trust an entire organization such as NIH rather than a particular person, Smith, from NIH. Next, they begin to look for a mismatch of credentials such as incorrect spelling, obscene or otherwise unusual pictures that do not match the expected content, etc. Finally they begin to evaluate the actual content, which they break down into bits of information, based around sentences, looking for facts they can use in their analysis.

When information is passed to the analysts outside of the typical QA&E model, they again begin a similar process. They evaluate the credentials of the individual passing the document to them as well as looking at the provenance (where the information came from, all the way back to source material). If the material had come from another analyst, the first measure used is that analyst's reputation. Even if an analyst is technically excellent, there will still be a need to check for bias that the analyst may be known or expected to embody with respect to the information area. There may even be a need to evaluate with segmentation (i.e., accept certain facts, but not others).

Although none of the analysts were able to name a system they currently used that could provide explanations for the results that were returned, they were able, through our 'Wizard of Oz' examples, to suggest some features they believed would be essential in such systems. Assuming an intelligent system that could reason over massive data and provide facts to specific queries (e.g., "where might PersonX be on September 19th, 2005?"), we investigated with our analysts how they might evaluate the answers provided.

Initially, they expect to see a basic answer for the query submitted. This could be as simple as a 'yes', 'maybe' or 'no' but some high-level, supportive material should also be provided or be available for answers to follow-up questions. This could be performed by presenting partial sentences, or an outline view of the underlying explanation. Based on the way the analytical mind works, our analysts made it clear that they would almost always ask for some supportive material. What comes next should be additional levels of explanation and reasoning support until we eventually reach the source material. For example, an initial answer to the question above could be "Springfield, VA". The next level of explanation could state knowledge of a train ticket reservation in the subject's name. Another level could present the raw record of the reservation from the train company.

If there are alternative explanations, the analysts indicated it was essential that they be shown within the same context. The most likely, or strongest explanation should be given due precedence. Most likely and strongest may be open to interpretation so

explanations of how likeliness or strength of explanation is determined should also be available. Additionally, analysts should have the option of exploring the evidence space and modifying assumptions. Thus, strongest explanations may need to be re-determined as a result of analyst exploration. Analysts desired the ability to have access to all the alternative explanations as well as the ability to explore and modify the evidence and assumption space.

Assumptions are an essential part of all analysis and an important element in evidentiary reasoning. Without using assumptions (a skill at which the human mind excels) even the most trivial problems can become intractable. Our analysts evaluated assumptions in regards to a number of dimensions, including whether or not the assumption is still valid, is it reasonable, does it describe a typical situation, etc. For example, we might have an assumption that states if a telephone is registered to PersonX and a call is made from that telephone to another number, then PersonX made that call. While understanding how important assumptions are, the analysts made a point that assumptions (especially those that are not universally shared) can cause problems in the analytical tradecraft, and that reasoners that use such assumptions should be transparent, presenting (in an understandable fashion) the logic behind their decisions.

4. DISCUSSION

While our interviews included a limited number of analysts, we have also gathered less formal input from other analysts through our participation in ARDA and other government programs. From those discussions, we offer the following points of agreement concerning requirements for next generation question answering and explanation systems.

a. Meta information provides valuable information. All of the analysts considered meta information to be potentially critical in the evaluation of source reliability. Author, author organization, citations of author, and date were all critically important pieces of data.

b. Multiple presentation strategies are useful. There seem to be as many preferences for presentation format as there are analysts. It does seem clear that some analysts work best with natural language presentations, some with graphs, some with formal representations, some with summaries, etc. One point of consensus is that many different presentation strategies are required.

c. Follow-up question support is critical. The analyst mindset and the requirements of the job both seem to lead to environments that provide extensive support for obtaining more granularity in support of any answer. Follow-up questions that are system generated (and thus do not require analyst knowledge of system syntax) were also valued.

d. Analytical techniques vary widely. Since analysts seem to have multiple strategies for obtaining and evaluating answers, question answering and explanation systems will need to support many models for obtaining and explaining answers.

e. Source trace-back is critical. All of the analysts agreed that some understanding of where the raw data was coming from, and in particular if it could be traced to the same source was critical. One thing that was common across all of the analysts was the need to determine if they might have a single source for what appears to be a diversity of sources reporting the same thing. They all increased their trust levels in information when it appeared to come from multiple reliable sources instead of just one.

f. Citations stating the author of source documents are one of the most important indicators of reliability of a source. The analysts all wanted to know who cited a fact before they were willing to consider it to be a highly reliable statement.

5. SUMMARY

In this paper, we have summarized our findings as we have collected requirements for next generation knowledge worker question answering and explanation systems. We have gathered the requirements using our implemented KANI system for intelligence analysts, focusing on the explanation component implementation and design. Our claim is that our implemented system provides a broad and reusable infrastructure that supports explanation in distributed analytic conditions. It embodies a direct implementation of the metadata provenance requirements that we gathered. It also provides an extensible foundation for including additional explanation presentation and interaction modalities. In particular, it facilitates information sharing, credibility assessment, increased trust, and collaboration. One thing the current implementation does not support is a special mode for citation summarization however this is under design.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the hard work and dedication of their teams at both Battelle and Stanford University, and their colleagues at IBM Watson. In particular they'd like to thank Richard Fikes, Paulo Pinheiro da Silva and Cynthia Chang (Stanford University), Bill Murdock (IBM Watson), and Alan Chappell, Liam McGrath, and Alex Donaldson (Battelle). They'd also like to thank the Battelle analysts that took part in this study for their time, diligence, and professionalism. This work is supported by the Advanced Research and Development Activity under the Novel Intelligence from Massive Data (NIMD) program. PNWD-SA-7220.

7. REFERENCES

[1] Shortliffe, E., Computer-based Medical Consultations: MYCIN. Elsevier, 1976.

[2] Swartout, W., Paris, C., and Moore, J., Explanations in Knowledge Systems: Design for Explainable Expert Systems. IEEE Intelligent Systems. June 1991 (Vol. 6, No. 3), pp. 58-64.

[3] McGuinness, D. L. and Pinheiro da Silva, P. Explaining Answers from the Semantic Web: The Inference Web Approach. Web Semantics: Science, Services and Agents on the World Wide Web Special issue: International Semantic Web Conference 2003 - Edited

by K.Sycara and J.Mylopoulis. Volume 1, Issue 4. Journal published Fall, 2004.

[4] Beyer H. and Holtzblatt, K. Contextual Design: Defining Customer-Centered Systems, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1998.