SUMMARY

The exponential increase in new knowledge that characterizes our modern information age precludes depending solely on individual effort to keep up with new information. We must develop new ways of keeping up, and we must develop them quickly. The semantic web offers a promise that we can keep up by allowing software agents to roam in cyberspace on our behalf, where they can gather information of interest and synergistically assist us in decision making. This dream, however, relies on agents being able to find and manipulate useful information, which, in turn, relies on having an abundance of ontologically described repositories. Furthermore, users need an effective way to query the semantic web, but any burden we place on everyday users to learn a query language is unlikely to garner sufficient user support and interest. If we want everyday users to take advantage of the semantic web, we must devise ways to transform existing, traditional web pages into semantic-web pages, and we must provide simple and unrestricted interfaces for processing user queries.

We propose using information extraction ontologies to answer these challenges. We show how ontology-based, data-extraction techniques can (1) automatically generate semantic annotations for traditional web pages and (2) support free-form, textual queries of semantically annotated sources. Achieving these objectives, however, depends on additional research in three fundamental support areas: (1) enhancement of automatic web-page annotation, (2) semi-automatic ontology creation, and (3) improved free-form query translation. Moreover, achieving these objectives requires the construction of simple, but exciting semantic-web showcases. These showcases must be able to attract everyday users. This can only happen if the showcases are easy to use and provide real benefit to their users.

Intellectual Merit. Building the semantic web is a grand challenge, and many researchers are contributing to its development. In contrast to existing approaches, we propose to advance the state of the art by developing technology to view the existing web through a semantic lens. We base our proposal on proven information-extraction techniques that build on our earlier NSF-funded work. We propose to add, however, three new fundamental research results: (1) a two-phase information extractor to improve extraction accuracy, (2) a by-example semi-automatic ontology learner to reduce the effort required to construct extraction ontologies, and (3) a three-pronged query parser to improve free-form query translation. Our contributions can (1) help pave the way for transforming portions of the existing, traditional web into the semantic web, (2) provide a way for ordinary users to interact with the semantic web without having to learn a specialized language and without being constrained by forms or other semi-structured constraints, and (3) enable software agents to act in behalf of their owners to search for information of interest.

Broader Impact. This project has the potential to make a real impact on society. The problem is significant, and a solid resolution will be a major advance in web technology, enabling ordinary users to reap the benefits of the semantic web. Proposed showcases—(1) for proactively connecting potential blogging partners and (2) for cooperatively building a distributed family tree of interconnected genealogical information—can themselves have an impact on bloggers and genealogical enthusiasts. We are also poised to make an impact on a number of students. Our research team is multi-disciplinary, representing computer science, e-business, and computational linguistics. Our student research team is geographically diverse and includes a significant group of underrepresented students. We plan to continue to maintain both an interdisciplinary and a diverse community of scholars. As we have done in the past, we will continue to publish our results and research artifacts on our web site and in peer-reviewed journals and conference proceedings.