SUMMARY

The current web is a web of linked pages. Frustrated users search for facts by guessing which keywords or keyword phrases might lead them to pages where they can find facts. Further, when needing to gather and organize facts from multiple sites, users are left without support and must manually organize facts of interest for further consideration. Can we make it possible for users to search directly for facts? And, once found, can we assist them in organizing found facts into bundles of knowledge for further consideration and to support decision making? As part of the organization, it may useful to retain links to extracted facts. Users could follow links to check original sources as they make decisions based on extracted facts. Sometimes these organized bundles of knowledge may be of interest to others. Could we facilitate sharing? Could we link sharable knowledge bundles together forming a web of knowledge? Even better, could this web of knowledge be superimposed over the current web of pages with provenance links tracing back to facts in original sources? And finally, with the web of knowledge organized for direct fact finding, could users query (rather than search) for facts and related facts in organized bundles of knowledge?

Answers to these questions call for distilling knowledge from the wealth of heterogeneous digital data on the web of pages, organizing the extracted data into queriable knowledge bundles, and linking facts in these knowledge bundles to each other and to the original facts. A computational alembic must turn raw symbols in web pages into facts in knowledge repositories, link these facts back to the original raw symbols and to related facts, and make this knowledge accessible via the web. To make this scalable, the computational alembic must run automatically (or nearly automatically); otherwise, the barrier to creating this web of knowledge will be insurmountable. We face three challenges: (1) automatic (or near automatic) creation of knowledge bundles, (2) automatic (or near automatic) annotation of web pages with respect to these knowledge bundles, and (3) simple, but accurate, query specification, usable without specialized training. Meeting these basic challenges would simplify knowledge-web content creation and access enough to enable this vision of a web of knowledge.

Based on prior work—that shows how to (1) automatically/semi-automatically reverse-engineer structured and semi-structured documents into extraction ontologies (NSF Grant 0414644); (2) automatically annotate web documents with extraction ontologies (NSF Grant 0083127); and (3) match free-form queries with extraction ontologies in order to generate structured queries over populated ontologies (NSF Grants 0083127, 0414644)—we propose to innovatively combine and extend all this work to meet these challenges. The end result is an architecture for creating knowledge bundles for decision making and for adding a knowledge-content layer over information-rich pages on the web that allows users to issue free-form queries for facts. Moreover, with provenance references linked directly to fact-origination statements, users can check retrieved facts in original source documents.

Intellectual Merit. The research is intended to provide an answer to the question about how to turn lexical and syntactic symbols into semantic knowledge and ultimately into a web of knowledge. The research is also intended to provide a way for untrained users to directly query for facts in fact-filled knowledge bundles and to enable provenance trace-back to fact sources. We plan to evaluate the accuracy of extracting and organizing facts from the web and of user accessibility over governmental, scientific, retail, and historical data.

Broader Impact. The research work has the potential to help people: (1) harvest and organize facts from the wealth of digital data on the current web; (2) harness and manage community knowledge; and (3) make facts on the web easily searchable by the general public. Educationally, we also intend to promote training and cross-fertilization among academic disciplines. Our research team is housed in three departments (Computer Science, Information Systems, and Linguistics). Further, as we have in the past, we will actively seek participation of underrepresented groups and demographic diversity.