SWWS Position Paper

Bryan Pelz, CEO, Fetch Technologies (pelz@fetch.com)
Dr. Steven Minton, CTO, Fetch Technologies (minton@fetch.com)

Today, the Internet is made up of countless distributed and autonomous sources that contain valuable information, but present it for human consumption. Humans get by with only informal conventions to facilitate communication, and normally do not require detailed, pre-specified standards for information exchange. In contrast, machines must communicate in a fashion that conforms to syntactic and semantic standards that have been carefully worked out in advance.

Fetch Technologies has developed a set of tools for accurately and reliably extracting data from websites and transforming it into a structured data format, such as XML. In the process, the data can be normalized and aligned to an arbitrary ontology, making it available to software agents. These tools provide a bridge between the Human Web and the Semantic Web. Today, this bridge is useful in providing a critical mass of information sources to the Semantic Web. In the future, these tools will allow agents access to a richer, fuller information context.

Key to these tools is the use of machine learning techniques to access, integrate, and transact with Human Web sources. This provides the scalability necessary to learn highly accurate extraction rules, to verify wrapper functioning, to automatically adapt to website changes, and to integrate disparate extracted data.Ý Furthermore, because our machine learning technology enables software agents to be trained to automatically recognize and extract semi-structured content, it provides a practical means for bootstrapping the Semantic Web. We believe that this type of bootstrapping is necessary to establish the critical mass required for broad adoption of the Semantic Web.