mation extraction system is for it to repair itself when an information extraction error occurs [7]. The problem of wrapper repair and maintenance is only beginning to be addressed by researchers (for example, [2, 9]); and has not been addressed for data extraction systems that utilize domain ontologies instead of position-based extraction rules.

In a system capable of wrapper recovery, the wrapper processor triggers a recovery routine when an error is detected during data extraction. This recovery routine attempts to repair the wrapper and resume the extraction process. Wrapper recovery and repair consists of two steps. First, the recovery routine must attempt to locate the target data within the revised page structure. If successful, the extraction rules must then be regenerated to match the new page format [2, 10]. If the extraction recovery is not successful, or the wrapper cannot be repaired automatically, the system should generate an error message so that a human can assist in the recovery process.

The processes involved in adaptive information extraction are discussed here in the context of an Amorphic Web information extraction system prototype.

An Integrated Web Information Extraction Solution

To illustrate the processes involved in Web information extraction, an information extraction system that combines position-based extraction, ontology-based extraction, and wrapper recovery was created. The Amorphic system can locate Web data of interest based on domain knowledge or page structure, can automatically generate a wrapper for an information source, and can detect when the structure of a Web-based resource has changed and act on this knowledge to locate the desired information. One key feature of the Amorphic system is that both the extraction rules and the output data are represented by XML documents. This approach increases modularity and flexibility by allowing the extraction rules to be easily updated (manually or automatically), and by allowing the retrieved data to either be converted to HTML for consumption by a human or returned in a SOAP envelope as part of a Web Service. The current Amorphic prototype represents a cost-effective approach to developing large-scale adaptable information extraction systems for a variety of domains. The Amorphic system, shown in Figure 2, consists of several modules:

- The *form/query processor* creates a user query by parsing the site's search form, combining the user query with the site's form elements, and sending the resulting search parameters to obtain the HTML search result pages.
- The *data extraction manager* examines the page structure and determines how best to parse the site. This module analyzes the content of the HTML page, and constructs extraction rules using the domain knowledge. The extraction rules are used to locate data of interest (tokens) within the HTML page.
- The *data extractor* pulls the specific data from the HTML pages.
- The *wrapper recovery system* is invoked when the Amorphic system cannot locate tokens within the Web pages.

A prototype Amorphic information extraction system has been implemented using Java.

Data Preprocessing. The HTML page undergoes several preprocessing steps before the data extraction is performed. First, a document is retrieved from the Web. The document is then processed using a HTML parser to obtain a representation of the Web page's structure. As HTML pages are composed of tags and text enclosed by tags, it is possible to represent a HTML page's layout by a tree of nested HTML tags that follows the Document Object Model (DOM). The parsing process separates HTML tags, attributes, and content. The Amorphic HTML parser uses the DOM parse-tree to create a *location-key* to identify the *content-text* found in the Web page. The location-key is a path expression that defines the set of nested tags that the content-text resides within [3]. Once the parsing process is complete, the Amorphic system examines the page structure to determine how

The current Amorphic prototype represents a COST-EFFECTIVE APPROACH to developing large-scale adaptable information extraction systems for a variety of domains.