the manual wrapper generation process [8]. The wrapper induction method automatically builds a wrapper by learning from sample pages.

Currently there are two principal methods for identifying interesting data within Web pages: ontology-based extraction and position-based extraction.

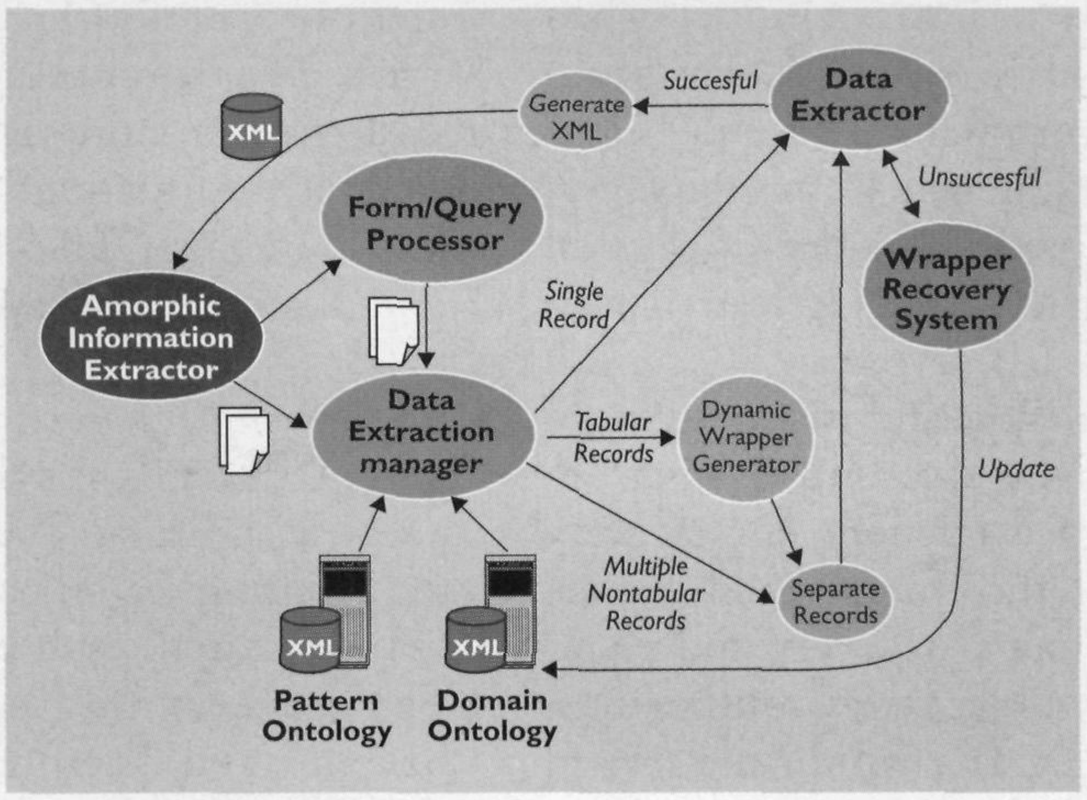

Figure 2. The Amorphic architecture.

Ontology-based Extraction. Ontology-based information extraction tools feature many of the properties desired for an adaptive Web information extraction system. An ontology-based tool uses domain knowledge to describe data. This includes relationships, lexical appearance, and context keywords. Wrappers generated using domain ontologies are inherently resilient (that is, they continue to work properly even if the formating features of the source pages change) and general, (they work for pages from many distinct sources belonging to a specific application domain) [4].

However, ontology-based tools require that the data be fully described using page-independent features. This means the data must either have unique characteristics or be labeled using context keywords. Unfortunately, all interesting Web data does not necessarily meet these requirements. Some data is freeform and cannot be identified using a specific lexical pattern and also is not labeled. This type of data can only be extracted using its specific location in the HTML page.

Position-based Extraction relies on inherent structural features of HTML documents to accomplish data extraction. Under a position-based extraction system, a HTML document is fed to a HTML parser that constructs a parsing tree that reflects its HTML tag hierarchy. Extraction rules are written to locate data based on the parse-tree hierarchy. If a collection of items is to be retrieved (as from a search results page), a regular expression is constructed to allow multiple items to be retrieved for a hierarchical pattern.

Position-based extraction lacks the resilience of ontology-based extraction. When there are changes to the structure of the target Web pages, it frequently fails. However, it does guarantee a high accuracy of information extraction, with precision and recall being at least 98% [2]. In addition, it is possible to use wrapper induction to create position-based wrappers based on a sample of regularly formatted Web pages. This can greatly speed the development and update of position-based wrappers [11]. Thus, position-based extraction can be appropriate when the data to be extracted can only be identified based on its location within a Web page and not on domain information.

Wrapper Recovery and Repair. The Web is a dynamic medium, and, as such, Web pages are frequently altered in structure and appearance. These changes are made by ISPs to offer additional content and functionality, increase ease of use, or make the Web page more attractive to new users. When a Web page's structure is changed, a wrapper can fail to find keywords or path expressions in the page and thus cannot complete the information extraction. In most information extraction systems, once a wrapper fails it must be manually recreated to conform to the new page structure, which slows the recovery process [1].

An important characteristic for an adaptive infor-

An effective Web information extraction system must interpret a wide variety of HTML pages and ADAPT TO CHANGES WITHOUT BREAKING.