

ExWrap: Semi-Automatic Wrapper Generation by Example

Bethina Schmitt

Michael Christoffel

Jürgen Schneider

Institute for Program Structures and Data Organization (IPD), University of Karlsruhe
Am Fasanengarten 5
D-76128 Karlsruhe

{ schmitt, christof, schneidj } @ ipd.uka.de

1. MOTIVATION

Within the WWW there are lots of different information retrieval services like search engines, news archives, product catalogs, or literature services. In order to support the user, meta search systems provide great benefits and synergies: For instance, a user query can be evaluated on a larger set of documents and by applying duplicate detection meta search systems can both improve the quality of the results and reveal different purchase options for the same document or product. Within the UniCats project [1] we develop a meta search system based on digital library services like bookstores, library OPACs, or archives of research papers.

Meta systems need a wrapping component that provides mappings between the different query and result formats of the underlying retrieval systems and the internal query and result representation. So, wrappers overcome syntactical and semantical heterogeneity. Actually, generating and maintaining wrappers is quite laborious and time-consuming, especially when they rely on the public web interfaces of the services. Thus, the challenge is to develop a mechanism for a fast wrapper generation and to design this process of generation as simple as possible because not only programmers but also librarians or even users should be able to generate wrappers in order to create useful meta search systems.

2. IDEA AND DESIGN OF EXWRAP

The ExWrap toolkit meets these challenges by a "Wrapping by Example" approach: A user conducts a sample search within the retrieval service that he wants to generate a wrapper for. And while formulating his search terms and browsing through the results the user marks the pieces of information that the wrapper should extract automatically later on.

Within the demonstration, we present our ExWrap toolkit [3] and show how to generate a wrapper for a typical online bookstore like Amazon – fairly quick and without the need of any expert knowledge. To issue the sample query, ExWrap supports navigation through HTML pages until the user has reached the page with the initial search form. ExWrap automatically extracts all available parameters and values, so the user can easily insert his search terms. Afterwards, the user can study the results (see Figure 1). ExWrap offers three kinds of views: a DOM-tree representation of the HTML code, a text-only view, and a typical browser view (①-③). The user can mark and name the interesting pieces of information, e.g. title, author, year, price, time of

delivery, ISBN, summary, ... Therefore, the user can navigate through the different levels of result pages. In Figure 1 (④) the user has already specified three attributes on the first level (root) and two attributes on the second level (details). Right now, he is going to define an ISBN attribute on the detail level.

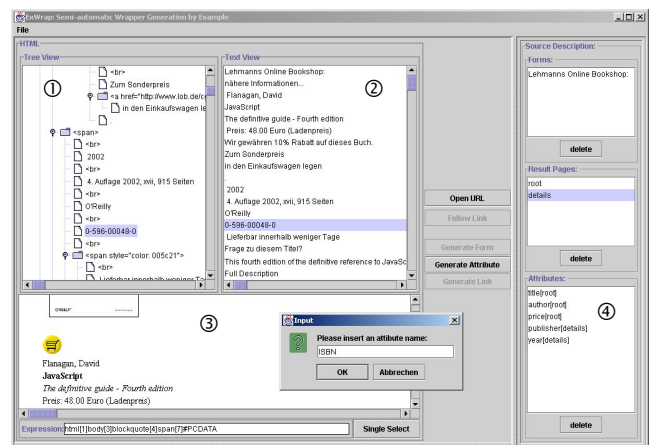


Figure 1. Semi-Automatic Wrapper Generation with the ExWrap Toolkit for a Typical Online Bookstore

Actually, the result of this sample search is not a wrapper program or any java code but a specific source description file for the queried retrieval service, which is stored in XML format. Our UniCats wrapper [2] is designed as a combination of a "generic" wrapper component together with a source description file which contains the specific properties of the retrieval service.

For further information about our wrappers, wrapper generation or other parts of the UniCats architecture, please visit our project homepage at <http://www.unicats.de>.

3. REFERENCES

- [1] Christoffel, M., Nimis, J., Pulkowski, S., Schmitt, B., Lockemann, P., 2000. An Infrastructure for an Electronic Market of Scientific Literature. In: Proc. of the 4th IEEE Intl. Baltic Workshop on DB&IS, Vilnius, 155-166.
- [2] Pulkowski, S., 2000. Intelligent Wrapping of Information Sources: Getting Ready for the Electronic Market. In: Proc. of the 10th VALA Conference, Melbourne, 113-124.
- [3] Christoffel, M., Schmitt, B., Schneider, J., 2002. Semi-automatic Wrapper Generation and Adaption: Living with Heterogeneity in a Market Environment. In: Proc. of the 4th ICEIS, Ciudad Real, 65-72.

