Journal of Universal Computer Science, vol. 14, no. 11 (2008), 1808-1810
submitted: 2/5/08, accepted: 25/5/08, appeared: 1/6/08 © J.UCS

Wrapping Web Data Islands

J.UCS Special Issue

Rafael Corchuelo
(Universidad de Sevilla, Sevilla, Spain
corchu@us.es)

José L. Arjona
(Universidad de Huelva, Huelva, Spain
jose.arjona@dti.uhu.es)

David Ruiz
(Universidad de Sevilla, Sevilla, Spain
druiz@us.es)

EAI, EII, and ETL are the three key abbreviations regarding integration. EAI stands for Enterprise Information Integration and refers to a number of technologies and best practices that help engineers integrate business applications, either to keep their data synchronised or to devise new emerging functionality [Hohpe and Woolf, 2003]. EII stands for Enterprise Information Integration; in this case, the focus is on creating live views of the data a number of applications manipulate [Kambhampati and Knoblock, 2003]. The focus of ETL, which stands for Extract, Transform, and Load, is on off-line data views that are typically used to feed business intelligence processes [Silvers, 2008]. According to a recent report [Weiss, 2005], companies spend $5–20 on integration per dollar spent on devising and implementing new applications. This is the reason why EAI, EII, and ETL are a common hobbyhorse for chief information and technology officers.

Our focus regarding integration is on web sites that do not provide a programmatic interface, which are very common nowadays. Such sites are difficult to integrate into automated business processes, which is the reason why we refer to them as web data islands. The Web Services or the Semantic Web initiatives [Papazoglou, 2007] [Antoniou and van Harmelen, 2008] provide excellent technologies by means of which web sites can provide a programmatic interface or, at least, provide data that is structured according to an ontology. However, re-engineering a web data island to endow it with a web service or with semantic annotations is not generally feasible. This has motivated many researchers to work on wrappers, which implement programmatic interfaces to web data islands by emulating the interaction of a human user, i.e., they fill forms in, navigate through the resulting pages, select the most appropriate data pages, extract data from them, structure them according to an ontology, and verify that the results are valid. Thanks to wrappers, integrating web data islands into automated business processes has become a common practice [Chidlovskii et al., 2006].

The goal of this special issue was to report on the state of the art regarding wrappers. We think we have succeeded in this endeavour since we have selected six papers that report on novel systems and techniques that help engineers build wrappers, namely:

- Jim Blythe, Dipsy Kapoor, Craig A. Knoblock, Kristina Lerman, and Steven Minton are the authors of the first article. They report on a system to help users query web sites, extract and ontologise the data they provide, and create complex procedures to exploit these data by means of a simple natural language interface.
- The article by Paula Montoto, Alberto Pan, Juan Raposo, José Losada, Fernando Bellas, and Víctor Carneiro reports on a language that helps engineers devise and implement wrappers. They criticise the common query wrapper model whereby a wrapper gets a query as input and outputs a result set, and support the implementation of wrappers in which navigation depends on the results that are being retrieved or wrappers that can insert, delete or update information.
- The third article was written by Marcio Vidal, Altigran S. da Silva, Edleno S. de Moura, and João M.B. Cavalcanti. It reports on a novel technique that uses a structural criterion to crawl a web site for pages about a given topic. The experiments prove that the technique is effective enough and can sort out the difference between closely-related topics, e.g., films and actors.
- Lorenzo Blanco, Valter Crescenzi, and Paolo Merialdo also focus on structural crawling, and present a technique that helps classify the template from which a given page is generated. It has also proved to be very effective with closely-related topics.
- The fifth article reports on a technique that helps identify what the areas of interest of a web page are. It relies on a visual segmentation algorithm and helps information extractors work more efficiently and effectively. The article was contributed by Jinbeom Kang and Joongmin Choi.
- Dawn G. Gregg reports on the results of a series of experiments she has conducted to explore how resilient information extractors are.

Acknowledgements

We would like to thank our reviewers for their hard work and enthusiasm to produce this special issue, and the J.UCS staff for their help and understanding. Without a shadow of doubt, they were fundamental to producing this issue, and we would not like to miss this opportunity to express our gratitude. The work of the guest editors was supported by the IntegraWeb project (grants CICYT TIN2007-64119 and JA TIC-2602). Part of the budget comes from FEDER funds.

References

[Antoniou and van Harmelen, 2008] Antoniou, G. and van Harmelen, F. (2008). *A Semantic Web Primer*. The MIT Press, 2 edition.

[Chidlovskii et al., 2006] Chidlovskii, B., Roustant, B., and Brette, M. (2006). Documentum ECI self-repairing wrappers: performance analysis. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pages 708–717.

[Hohpe and Woolf, 2003] Hohpe, G. and Woolf, B. (2003). *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley.

[Kambhampati and Knoblock, 2003] Kambhampati, S. and Knoblock, C. (2003). Information integration on the web. *IEEE Intelligent Systems*, 18(5):14–15.

[Papazoglou, 2007] Papazoglou, M. (2007). *Web Services: Principles and Technology*. Prentice Hall.

[Silvers, 2008] Silvers, F. (2008). *Building and Maintaining a Data Warehouse*. Auerbach.

[Weiss, 2005] Weiss, J. (2005). Aligning relationships: Optimizing the value of strategic outsourcing. Technical report, IBM.