

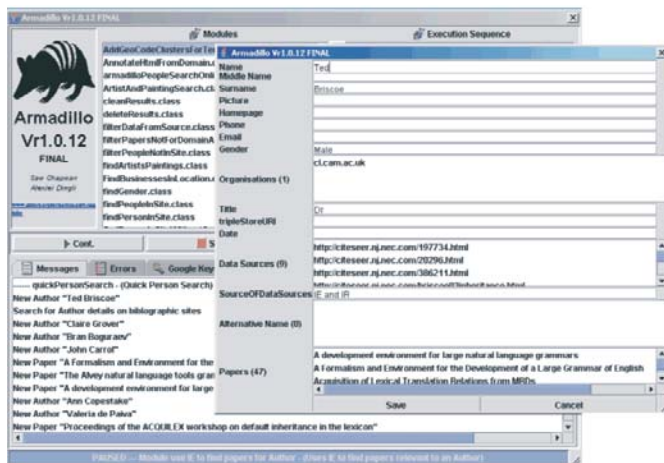
# Armadillo: Harvesting Information for the Semantic Web

Sam Chapman, Alexiei Dingli, Fabio Ciravegna  
Department of Computer Science, University of Sheffield  
Regent Court, 221 Portobello Street, Sheffield, S1 4DP, United Kingdom

N.Surname@dcs.shef.ac.uk

Armadillo [1] [2] is an automatic system for producing domain-specific Semantic Web oriented annotation on large repositories. It annotates by extracting information from different sources and integrating it into a knowledge base. Such base can then be used both to access the information directly (e.g. via a semantic web agent) and to annotate the pages where the information was identified. Armadillo is adaptive: it learns how to harvest information with minimal initial user intervention. The first step performed in the learning process is identifying a number of seed terms that are examples of information to be extracted. They can be provided by the user either via a small lexicon or via connection to a web service. These annotations are used to seed learning from different parts of the repository or even in the external world (e.g. on the Web). An agent spiders the available space and identifies places where such terms occur (documents, databases, etc.). Rules are induced to model the context in which these terms appear. Such rules are then used to extract other examples not contained in the initial lexicon but that appear in similar contexts. All new terms must be confirmed before they are accepted and used to re-seed learning. Multiple strategies are used for confirmation, e.g. a new piece of information is accepted if it is found in different (linguistic or semantic) contexts. Finally all the extracted information is integrated with the existing one in the knowledge base. The methodology relies on the inherent redundancy of large repositories, e.g. the Web or company-wide repositories. By inherent redundancy we mean the fact that information is frequently represented in different forms in large distributed resources.

The architecture is based on Semantic Web Services. Each service is associated to some parts of the ontology (e.g. a set of concepts and/or relations) and works in an independent way. Each service can use other services (including external ones) for performing some sub-tasks. For example a service recognising researchers names in university web sites will use a Named Entity Recognizer as a sub-service to recognise potential names (i.e. generic people's names) and confirm them as real researchers names (e.g. as opposed to secretaries'



names) using some internal strategies. An RDF repository is used to store the extracted information. It is also the medium of communication among the different agents. A development environment allows to define architectures for new applications. Porting to new applications does not require knowledge of IE. All the methods used tend to be domain independent and are based on generic strategies to be composed for the specific case at hand. The only domain dependent parts are: the initial lexicon, the ontology and the way the confirmation strategies are designed/composed. Armadillo has been tested in the domains of (1) discovering information about Computer Science departments and (2) extracting full lists of paintings, artists and their images from the Web.

In this demo we will illustrate the methodology, the way in which an application can be developed, show the system and some experimental results. We will conclude highlighting some future developments and related challenges.

## 1. REFERENCES

- [1] Fabio Ciravegna, Sam Chapman, Alexiei Dingli, , and Yorick Wilks. Learning to harvest information for the semantic web. In *Proceedings of the First European Semantic Web Symposium*, May 2004. Crete.
- [2] Fabio Ciravegna, Alexiei Dingli, David Guthrie, and Yorick Wilks. Integrating information to bootstrap information extraction from web sites. In *Proceedings of the IJCAI 2003 Workshop on Information Integration on the Web*, 2003. Acapulco, Mexico, August, 9-15.