Web Service Crawling And Annotation

Nathalie Steinmetz[1], Adam Funk[2], Maria Maleshkova[3],

[1] seekda GmbH, Grabenweg 68, 6020 Innsbruck, Austria
[2] University of Sheffield, 211 Portobello, Sheffield, S1 4DP, United Kingdom
[3] Knowledge Media Institute, The Open University, Milton Keynes MK7 6AA, United Kingdom

Abstract. The aim of this tutorial is to describe how semantic technologies can be employed for improving the large-scale discovery of Web Services. We concentrate on working with data dumps coming from a Web Service crawler that focuses only on service relevant parts of the Web, on building unique service objects out of the data and on semi-automatically annotating these services. Input to this tutorial comes from actual results of the FP7 European R&D projects Service-Finder and SOA4All.
The tutorial will show in detail which kind of data we will have to deal with, i.e. WSDL services, Web APIs (a.k.a. RESTful services) and related Web documents. We will address the topic of document deduplication and explain how to infer unique service objects from the data. In a next step we will show how to automatically extract information from Web documents (both structured and unstructured data), and use them for semantic enrichment of the services. The automatic annotations, exemplified with the use of GATE, consist of both category annotations and generic service information annotations, described by the means of specific service ontologies. We will introduce these ontologies and will detail the semantic annotation via hands-on sessions. In a last step we will show how to use MicroWSMO to semantically annotate RESTful services.

1 Tutorial Outline

We are today witnesses of an important change in the Web area: the Web is moving from a mere collection of pages to a collection of services. This goes hand in hand with the rise of Service-Oriented Architectures (SOA) as dominant computer paradigm. While SOAs, together with Web Service technologies, are widely seen as very promising fundament, the success of the new service orientation is depending on resolving still open challenges like, amongst others, Web Service discovery on a large scale. Generic search engines provide fast and easy access to existing Web pages; the two European projects Service-Finder and SOA4All work, amongst others, on solutions that allow a similar easy and scalable access to publicly available Web Services.
The goal of the tutorial is to introduce the approach that we use both within Service-Finder and SOA4All to locate Web Services on the Web and to semantically annotate them in a way that allows easy and scalable search. The tutorial will provide an overview of the data that results from running a Web Service crawler that concentrates on the service relevant parts of the Web. Participants will get access to this data that consists of Web Services (WSDL and RESTful services) and related documents. They will learn how to work with and analyze the data and how to build unique service objects out of it. The tutorial will provide both

theoretical and practical aspects of semantic annotation of Web Services and will familiarize the participants with the usage of GATE and MicroWSMO to produce (semi-)automatic service annotations.
The tutorial will be organzied as a full-day tutorial and will be structured as follows (not taking into account breaks):

Introduction / Basics - *45 min presentation.* The presentation will provide an overview of Web Service discovery in general and will introduce the idea of crawling the Web for services. It will show how lightweight semantic annotations can be used to enhance the search for services on a large scale.

Crawl Data. - *60 min interactive presentation.* During this session we familiarize the participants with the data that results from a Service Crawler. They will get access to real large data sets and will learn how to handle them. The data will be used in the later hands-on sessions when the participants learn how to provide semantic annotations for the services.

Service Ontologies. - *45 min presentation.* This presentation will provide an overview of the different ontologies that we use within Service-Finder and SOA4All to store all service-related meta-data. This meta-data encompasses both functional and non-functional service information. We will work with these ontologies in the following hands-on sessions.

Introduction to GATE. - *45 min interactive presentation and demonstration.* This presentation will provide an introduction to the GATE framework and architecture. The participants will be shown how to use GATE to semantically annotate WSDL descriptions and related documents, an approach that is followed within the Service-Finder project.

GATE Hands-on Session. - *60 min hands-on session.* During this session the participants will, under supervision of the tutorial presenters, apply the methods and tools introduced in the previous session to provide semantic annotations of services. They will learn how to categorize services and how to extract specific information from the documents.

Introduction to MicroWSMO. - *45 min interactive presentation and demonstration.* This presentation will provide an introduction to MicroWSMO and to the MicroWSMO editing tool that has been developed in the scope of SOA4All. The participants will be shown how the editor can be used to easily annotate RESTful services.

MicroWSMO Hands-on Session. - *60 min hands-on session.* During this session the participants will, under the supervision of the tutorial presenters, actively use the MicroWSMO editor that has been presented in the previous session. They will learn how to recognize important information in RESTful service documents and how to accordingly annotate them.

Questions and Answers. - *30 min discussion.* The tutorial is concluded with a final questions and answers session where the participants and the presenters can discuss open questions and wrap up the day.

This tutorial has relevance to the Future Internet Symposium in that it provides training in an area whose importance is perpetually growing: Web Services. Considering the importance that Web Service and SOA technologies will take in the future, training people in how to semantically enrich the services and thus enhance their easy location and accessibility is crucial for the long-lasting success of SOA.

2 Tutorial Format

The tutorial will start as a lecture, providing the necessary basic knowledge on Web Services, Web Service discovery, crawling and ontologies to the participants. Interactive lectures will familiarize the participants with crawled service data and with the tools that we will use to semantically annotate the services: GATE and the MicroWSMO editor. Two hands-on sessions will allow the participants to actively apply the tools under the supervision of the tutorial presenters. The tutorial will conclude with an questions and answers session and open discussion.
We have outlined the format of the single sessions in the tutorial outline description above.

3 Recording

We agree to any recording of the tutorial. If such recording takes place, we expect the recorded materials to be made available publicly after the conference, or the presenters should have access to the materials, and the right/license to make them public themselves.

4 Intended Audience

The tutorial is intended to all people interested in SOA and Web Services, and interested in how to enrich them with semantic annotations to improve their easy searchability. The participants do not need to have any programming knowledge, as they will learn how to apply tools and will not need to program.

5 Prerequisites

There are no strict prerequisites to the tutorial attendees regarding their former experiences, as we will start the tutorial with an introduction into all the relevant topics. Some knowledge or experience with Web Services and semantic annotation technologies will nevertheless be an advantage.

6 Presenter Requirements

All the tutorial attendees should have electrical power. Internet connection would be good, but is no absolute prerequisite.

7 Presenters Short Biography

Nathalie Steinmetz (nathalie.steinmetz@seekda.com, +43 1 2365084) is finishing her master in computer science at the University of Innsbruck at present. She works since 2005 in the area of Semantic Web and Web Services as junior researcher and project collaborator in national and international projects. Her master thesis is on Web Service Location and Nathalie has extensive experience in Web Service Discovery. Beginning of 2008 she joined seekda GmbH, a start-up from STI Innsbruck, where she works as researcher and project manager. Her main research interests are in technologies around the Semantic Web, Web Services, Service-Oriented Architectures and Search Engines, with a special focus on discovery and location of (Semantic) Web Services.
Nathalie is involved in both the Service-Finder and the SOA4All projects. She mainly works on the focused service crawling, trying to detect as many as possible services and related documents on the Web. The crawling encompasses both WSDL services and RESTful services. Out of the crawled service data unique service objects are built and some basic metadata is collected. Nathalie was as well involved in building some of the ontologies that are used within the two projects to store the service meta-data.

Adam Funk (a.funk@dcs.shef.ac.uk, +44 114 222 1959) received his PhD in computational linguistics from the University of Manchester in 2005, and worked at the National Centre for Text Mining in Manchester until he joined the NLP Research Group at the University of Sheffield in January 2006. At Sheffield, he has worked on several projects (in particular, SEKT, LIRICS, MUSING and Service-Finder) in the fields of semantic web development, implementation of linguistic annotation standards, information extraction, and business intelligence.
Adam is involved in the Service-Finder project, where he works on the automatic service annotation. He takes WSDL files and related documents that result from a Service Crawler and analyses them to generate semantic descriptions of the services and their providers. He employs information extraction and machine learning techniques with the GATE library and architecture to categorize each service under concepts from the *Service Category Ontology*, to extract information such as contact details and pricing information, and to classify documents by genre and level of interest.

Maria Maleshkova (m.maleshkova@open.ac.uk, +44 1908 653800) is a PhD student focusing on supporting and enhancing the creation of semantic annotations over Web Services. In particular, her work focuses on providing tool support for the semantic annotation of RESTful services with MicroWSMO and developing approaches for the automatic service annotation recommendation.
Maria is involved in the SOA4All project, where she focuses on the development of the provisioning platform. Some of her main tasks within the scope of this project include the development of an approach for the semi-automatic extraction of semantic descriptions of services, based on service-related documents and user feedback. In particularly focusing on, developing a methodology for the recognition of service properties in text-documents by extracting terms, which are characteristic for a particular service description. In addition, she works on the development of a user editor for WSMO Lite and MicroWSMO, which will provide the user with options to edit and improve the resulting semantic service annotations, by adding supplementary information and comments.