



STI · INNSBRUCK



Web Service Crawling And Annotation

Crawl Data

FIS 2009, Berlin, 1 September 2009

Nathalie Steinmetz - STI Innsbruck / seekda GmbH



-
- Crawl Output
 - Get Access To The Data

-
- **Crawl Output**
 - Get Access To The Data

- Archives in format '.arc' (Internet Archive Format)
 - size: ca. 100 MB
 - contains all documents (WSDLs, related documents, Web APIs) including their HTTP header
 - PDFs have been transformed to HTML
 - HTML pages have 'light' mark-up (some purely layout related mark-up is removed during the crawl)
 - All documents have an Archive header to start with

Example Archive Entry

http://abn.business.gov.au/(psb5wx45prykhuhvnl5qgaf)/downloads/UsingABNLookupwebservices.pdf
210.193.176.124 20090702104819 text/html 92725

HTTP/1.1 200 OK

X-seekda-Original-Content-Length: 441695

Content-Type: text/html

Last-Modified: Wed, 22 Apr 2009 22:00:42 GMT

Accept-Ranges: bytes

ETag: "0216bc795c3c91:7bf"

Server: Microsoft-IIS/6.0

X-Powered-By: ASP.NET

Date: Thu, 02 Jul 2009 12:48:19 GMT

Connection: close

X-seekda-Filter: pdf2html, lightenedHtmlMarkUp

X-seekda-New-Content-Length: 91715

```
<html><title>Using ABN Lookup web services</title><meta content="text/html; charset=UTF-8" ><meta  
content="pdftohtml 0.36" name="generator" ><meta content="Theresa Macgregor" name="author" ><meta  
content="2009-04-23T08:00:41+00:00" name="date" ><meta content="Revision: 6" name="subject" ></meta></  
meta></meta></meta></meta><body>
```

...

- RDF dump

- Information about services, providers, operations, related documents, Web API scores, location in the archives, etc.

- Example:

```
<http://seekda.com/providers/fraudlabs.com> a <http://www.service-finder.eu/ontologies/ServiceOntology#Provider> ;
```

```
  <http://www.service-finder.eu/ontologies/ServiceOntology#hasDomain> <http://fraudlabs.com> .
```

```
<http://fraudlabs.com> a <http://www.service-finder.eu/ontologies/ServiceOntology#Domain> .
```

```
<http://seekda.com/providers/fraudlabs.com/IP2ProxyWebService> a <http://www.service-finder.eu/ontologies/ServiceOntology#Service> .
```

```
<http://ws.fraudlabs.com/ip2proxywebservice.asmx?wsdl> a <http://www.service-finder.eu/ontologies/ServiceOntology#WSDLDocument> .
```

```
<http://seekda.com/providers/fraudlabs.com/IP2ProxyWebService> <http://www.service-finder.eu/ontologies/ServiceOntology#hasProvider> <http://seekda.com/providers/fraudlabs.com> ;
```

```
  <http://www.service-finder.eu/ontologies/ServiceOntology#associatedWithEffectiveTopLevelDomain> <http://fraudlabs.com> .
```

```
<http://seekda.com/crawl/documentAnnotations/8ba15cc3-5c81-404f-a358-b875b2a0d07c> a <http://www.service-finder.eu/ontologies/ServiceOntology#DocumentAnnotation> ;
```

```
  <http://www.service-finder.eu/ontologies/ServiceOntology#belongsToDocument> <http://ws.fraudlabs.com/ip2proxywebservice.asmx?wsdl> ;
```

```
  <http://www.service-finder.eu/ontologies/ServiceOntology#isAboutEntity> <http://seekda.com/providers/fraudlabs.com/IP2ProxyWebService>
```

-
- Crawl Output
 - **Get Access To The Data**

- seekda provides an endpoint where you can download the crawled Web Service data
 - RDF dump
 - Archives containing WSDLs, related documents and Web APIs
- How to get the data:
 - Register yourself on the seekda Web Service search engine (<http://seekda.com>)
 - Write an email to nathalie.steinmetz@seekda.com
 - Add the email you used to register at seekda.com
 - Sign a Web Data License
 - You are only allowed to use the data for research purposes
 - Download the Data

QUESTIONS?