Faster Smart Data Prototyping with the Self-Service Semantic Suite (S4)

Marin Dimitrov

Ontotext AD

116W 23rd Street, Suite 500, New York, USA

marin.dimitrov@ontotext.com

Summary

The Self-Service Semantic Suite (S4) provides an integrated platform for on-demand semantic data management. With S4 developers get instant access to various capabilities for text analytics, knowledge graphs and RDF graph database-as-a-service in the Cloud. By providing an easily and instantly accessible set of services, the S4 platform enables faster and cheaper prototyping of applications for Smart Data analytics.

Introduction

The goal of the Self-Service Semantic Suite[1] (S4) is to increase the speed and reduce the cost of building Smart Data analytics prototypes based on Semantic Technology. Startups often have limited resources for evaluating and prototyping with novel technologies and on premise hardware and licensing costs create additional barriers to entry. S4 provides a platform for text analytics, knowledge graphs and RDF graph database-as-a-service in the Cloud, so that developers can easily and instantly access such capabilities (Figure 1).

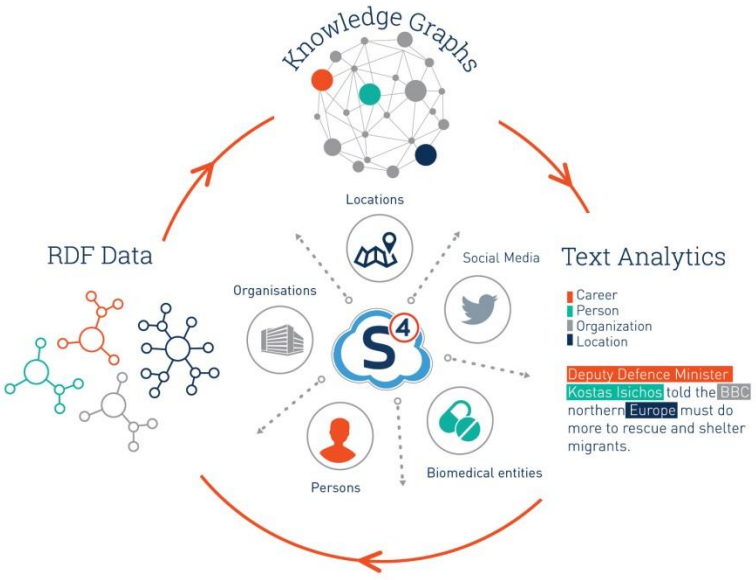

Figure 1 The Self-Service Semantic Suite (S4)

[1] http://s4.ontotext.com/

Text Analytics Services

The S4 platform provides various services for text analytics over unstructured content:

- *News analytics* – information extraction, disambiguation and entity linking to concepts and instances from the DBpedia, Wikidata and GeoNames knowledge graphs.
- *News classifier* – categorisation of news articles according to the 17 top-level categories of the IPTC Subject Reference System[2].
- *Biomedical analytics* – the service can recognize more than 130 biomedical entity types and semantically link them to a large-scale biomedical LOD knowledge base, LinkedLifeData[3].
- *Twitter analytics* – based on the TwitIE[4] open source microblog analysis pipeline, the service performs named entity recognition of various classes of entities found in tweets

Knowledge Graphs

S4 provides a reliable access to key datasets from the LOD cloud via the FactForge[5] semantic data warehouse: more than 5 billion LOD triples, describing 500 million entities from integrated and aligned datasets – such as DBpedia, Freebase/Wikidata, GeoNames, and MusicBrainz – are available to S4 developers. The text analytics services on the S4 platform also provide mappings to concepts and instances from the LOD datasets.

RDF Graph Database-as-a-Service

S4 provides a fully managed RDF graph database-as-a-service based on the GraphDB[6] RDF database. The fully managed RDF graph database provides a 24/7 access to private RDF databases and SPARQL endpoints within a multi-tenant model. Key operational aspects such as security, high availability and backups are handled by the S4 platform on behalf of the developers.

Tools for Developers

Even though access to all S4 platform capabilities is available via simple RESTful services, additional effort was put into developing various plugins, add-ons and SDKs that speed up the development and prototyping.

GATE & UIMA Plugins

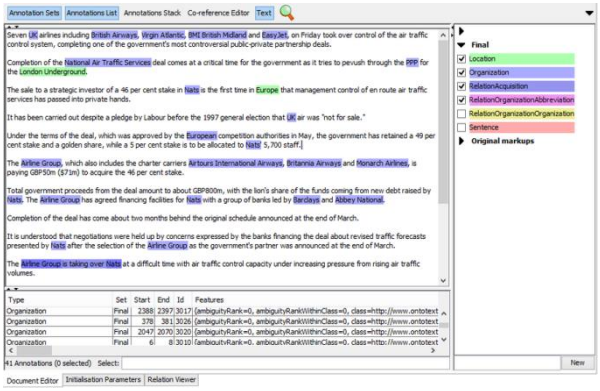

Figure 2 GATE plugin for S4

[2] http://www.iptc.org/std/NewsCodes/0.0/documentation/SRS-doc-Guidelines_3.pdf

[3] http://linkedlifedata.com/

[4] https://gate.ac.uk/wiki/twitie.html

[5] http://factforge.net/

[6] http://www.ontotext.com/products/ontotext-graphdb/

The plugin to the General Architecture for Text Engineering[7] (GATE) platform allows for GATE developers and language engineers to embed S4 text analytics services into complex text processing workflows and applications (Figure 2).

A similar plugin, as well as an SDK, is available for the Unstructured Information Management Architecture[8] (UIMA) text analytics platform (Figure 3).

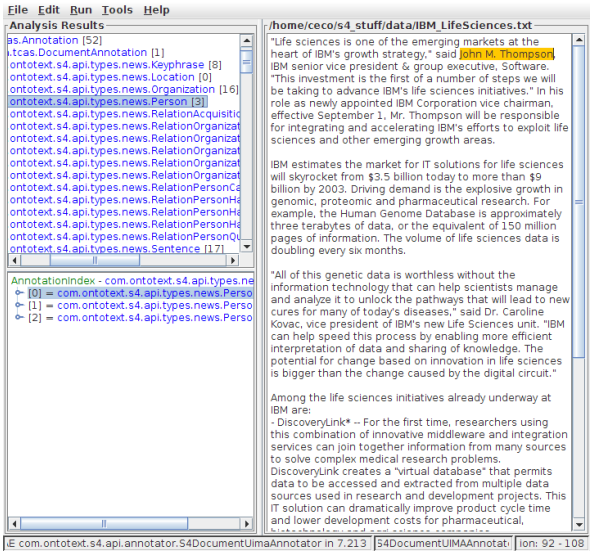

Figure 3 UIMA plugin for S4

Firefox & Chrome Plugins

The Firefox and Chrome browser plugins which allow that web page snippets be quickly annotated with S4 text analytics services directly from the browsers (Figure 4).

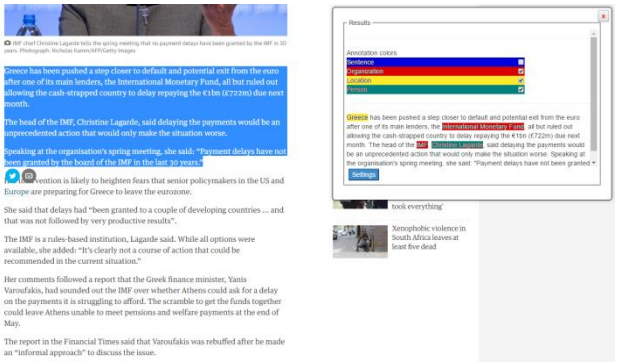

Figure 4 Chrome plugin for S4

SDKs

Java, C# and Python SDKs that provide developers with easy access to the S4 platform services in their programming language of choice.

All RESTful APIs have detailed Swagger[9] descriptors, so that additional SDKs for other languages can be easily generated by developers.

S4 in Use

S4 is available for free to developers (within certain quota limits) and it is already being applied in various use cases:

- More than 1,000,000 text documents are being processed per month by 3rd party applications
- More than 50 free RDF graph DBaaS instances are deployed and used, with several new instances deployed and operated on behalf of 3rd party developers each week
- The DBaaS part of S4 has been applied and tested in practice within 2 EU research projects – DaPaaS[10] and ProDataMarket[11] – in order to provide large-scale Linked Data hosting capabilities for Open Data publishers. The elastic infrastructure of the RDF DBaaS significantly reduces the operational cost and complexity for hosting a large number of 3rd party Open Data sets.

[7] http://gate.ac.uk/

[8] https://uima.apache.org/

[9] http://swagger.io/

[10] http://dapaas.eu/

[11] http://prodatamarket.eu/

- The scalable text analytics infrastructure of S4 is being applied and tested in KConnect[12] in order to quickly bootstrap a platform for EHR analytics with custom biomedical text analytics services – researchers were able to get a working infrastructure for semantic analytics within the first few weeks of the project, instead of spending time & effort on integrating, deploying and maintaining such an infrastructure.
- Several internal POC prototypes have been built by Ontotext on top of S4, in order to quickly showcase Smart Data analytics capabilities to prospective customers, at a significantly faster rate than the historical average.

We believe that such a platform is useful to researchers and developers and reduces the time, cost and complexity of building Smart Data prototypes, which are among the limiting factors for the wider adoption of Semantic Technologies. While the S4 platform is based on proprietary technology by Ontotext, we are committed to continue to provide large free usage quotas to 3rd party developers, researchers and applications.

Experiment more, experiment faster!

[12] http://kconnect.eu/