# AnnoMarket – Multilingual Text Analytics at Scale on the Cloud

Marin Dimitrov[1], Petar Kostov[1], Hamish Cunningham[2], Ian Roberts[2], Philippe Rigaux[3], Helen Lippell[4]

[1] Ontotext AD, Bulgaria
{marin.dimitrov, petar.kostov}@ontotext.com
[2] Department of Computer Science, University of Sheffield, UK
{h.cunningham, i.roberts}@ dcs.shef.ac.uk
[3] Internet Memory Research SAS, France
philippe.rigaux@internetmemory.net
[4] The Press Association Ltd, UK
helen.lippell@pressassociation.com

**Abstract.** AnnoMarket is an FP7 funded project providing an open platform for cloud-based text analytics services and language resources acquisition. Providers of text analytics services and language resources can deploy and monetize their components via the AnnoMarket platform, while users can utilize such available resources in multiple languages and in various domains in an on-demand, pay-as-you-go manner.

**Keywords:** text mining, cloud computing, software-as-a-service, linked data

## 1 Introduction

AnnoMarket[1] is an FP7[2] project that aims to revolutionise the text analytics market, by delivering an open marketplace for pay-as-you-go, cloud-based text mining resources and services, in multiple languages. The current services available on the AnnoMarket marketplace[3] are applicable to a wide set of business cases, e.g. large-volume multi-lingual information management, business intelligence, social media monitoring, customer relations management.

The Software-as-a-Service delivery model adopted by AnnoMarket reduces the complexity of deployment, maintenance, customisation and sharing of text processing resources and services by SMEs and developers. The marketplace currently provides various services for multilingual information extraction and semantic annotation, sentiment detection, as well as multilingual web corpora, public language resources and LOD datasets.

---

1 https://annomarket.eu/

2 The AnnoMarket project is funded by the European Commission under the 7th Framework Programme, Project No. 296322

3 https://annomarket.com/

adfa, p. 1, 2011.

## 2 AnnoMarket Platform

The AnnoMarket platform is comprised of components for language resource acquisition, scalable and elastic text mining over large volumes of data, usage monitoring and quota enforcement, as well as billing and online payments.

### 2.1 Language Resource Acquisition

The language resource acquisition component of the AnnoMarket platform is based on the large scale web crawling infrastructure providing user-defined, on-demand crawls at a large scale. In addition to corpora crawled on-demand from the web, the language resource acquisition component provides integration with the Common Crawl[4] dataset.

### 2.2 Multilingual Text Mining Services

Various multilingual text mining services are currently deployed and ready to use via the AnnoMarket platform. The current set of services includes more than 30 different text processing pipelines covering 17 languages.

The text processing pipelines vary from low level ones (stemmers, part-of-speech taggers, noun phrase chunkers and parsers), to general purpose pipelines (named entity recognisers) and domain specific pipelines (for the bio-medical domain, news publishing domain, or sentiment analysis over social media).

### 2.3 Marketplace

The marketplace provides an eShop[5] where customers can explore the catalogue of available text analytics services, language resources and datasets as well as additional processing resources available on-demand via the platform (e.g. an LOD server hosting Freebase, DBpedia and GeoNames; indexing servers, etc.). All products deployed on the marketplace provide information about their functionality and the associated usage and pricing terms.

### 2.4 Cloud Platform

The AnnoMarket platform is currently deployed on the Amazon Web Services[6] public cloud and it utilizes various cloud services for storage, computing, scalability, and a design for a multi-datacenter deployment for improved availability.

---

4 http://commoncrawl.org/

5 https://annomarket.com/shopfront

6 https://aws.amazon.com/