1. Publishable Summary

Many businesses recognise the importance of text analysis to glean useful knowledge from the huge volumes of digital content produced on a daily basis. While many companies still employ manual analysis, the manual information management approach is already extremely expensive and is becoming prohibitively so. As a result, more and more companies are turning towards automated text processing – a series of studies published in 2009 - 2013[1] of the intelligent text processing market estimated the overall market size at $1 billion by the end of 2011, and growing by 25% annually.

The current offerings in the text analysis market are typically either standalone software solutions (proprietary or open-source), which require users to invest significant time and money in their own server hardware to operate, or web-based services that process documents one at a time on a service provider's hardware, which is slow, potentially expensive if your usage exceeds the free limits, and difficult to customise for new requirements. Additionally, existing services support only a small selection of languages – coverage is good for languages like English, French and Spanish, but there are few options for users who need to process newer EU languages such as Bulgarian or non-EU languages such as Russian. Furthermore, a single provider will typically specialise in just one or a few languages, so a user will have to integrate several different and incompatible solutions to handle a broader range.

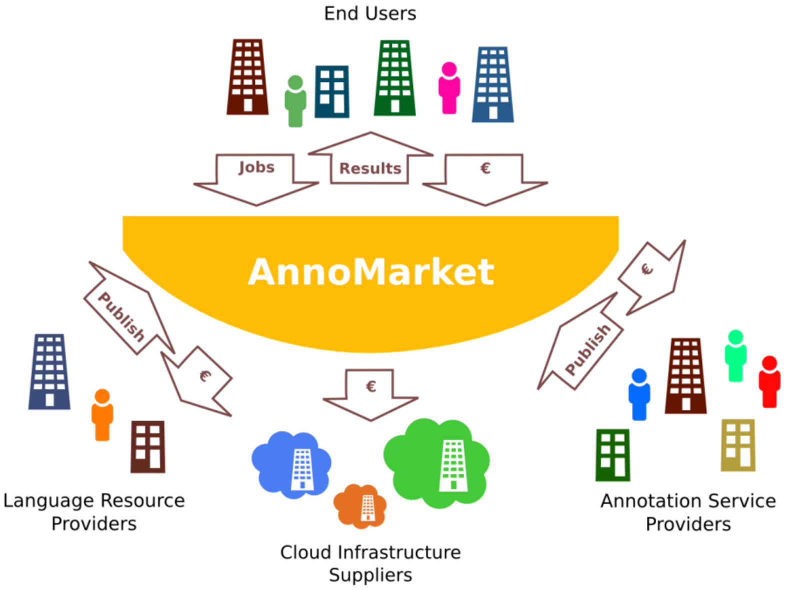
The AnnoMarket project aims to revolutionise the text annotation market, by delivering an affordable, open market place for pay-as-you-go, cloud-based text mining resources and services, in multiple languages. The project intends to develop:

- an open marketplace platform to support deployment of language resources and processing tools on the cloud, on a pay-per-use basis. Users will be able to run text analysis tools over their own data or over resources such as web crawls provided through the platform
- multilingual resources and tools for information extraction, sentiment analysis and topic detection and tracking, integrating Linked Open Data resources and made available via the marketplace platform alongside third-party offerings
- generic corpora of billions of words in multiple European languages available to research and commercial users

The key emphasis in the AnnoMarket platform is *openness* – suppliers of text analysis tools and resources will be able to use the platform to make their tools available for sale, allowing them to monetise their services without needing to invest in their own hardware or the expertise to handle the potentially unreliable public cloud infrastructure. AnnoMarket will handle the infrastructure, billing and payments, leaving the text analytics developer to focus on their own tools. Since the platform will provide a common API for all the disparate tools that are published in the marketplace, it will be easy for users to chain together different tools and resources for their purposes, unlike the current situation where every service has its own (incompatible) API and the user must do the integration work themselves.

The various stakeholder groups that are participating in the AnnoMarket concept are illustrated in the figure below.

[1] S. Grimes, AltaPlana. Text Analytics 2009: User Pesrectives on Solutions and Providers, 2009
S. Grimes, AltaPlana. Text/Content Analytics 2011: User Perspectives on Solutions and Providers
S. Grimes, AltaPlana. Text-Analytics Demand Approaches $1 Billion, 2011
S. Grimes, AltaPlana. Trend Report: Text Analytics in 2013.


End Users
Jobs
Results
€
AnnoMarket
Publish
€
Publish
€
€
Language Resource Providers
Cloud Infrastructure Suppliers
Annotation Service Providers

Pricing will be transparent, e.g. pricing per thousands of API calls, GB of input data for the extraction services, and thousands of URLs crawled. In this way, users will be able to estimate in advance the likely overall costs and monitor them at runtime.

The project has just completed its first year, during which the consortium has been busy working in three main directions: technical work, business development planning, and user experience and engagement.

One main element in our **technical efforts** has been the development of the AnnoMarket platform. Work has proceeded in parallel on the back-end services and the front-end portal. The back-end work has reached a stage where sufficient features are now available to provide a useful service. This will form the base of the first platform prototype which we are hoping to open to early adopters in late 2013. The front-end work has concentrated on providing a polished, unified, friendly, and attractive user experience. The results of this effort will debut with the second platform prototype, which will also be the first public release.

Another element in our technical work was the collection and production of text mining pipelines that will form the initial offering when the platform launches to the public. We are aiming to provide a wide enough range to achieve critical mass and make the platform an attractive venue for end users and service providers alike. At the end of the first project year there are 35 pipelines already prepared, with 4 additional candidate pipelines being considered for inclusion. The entire set covers 15 different languages, with an additional one (Bulgarian) being planned for inclusion. The types of analysis range from simple stemming to syntactic parsing and named entity recognition. Most pipelines are general purpose, while others are specialised on particular domains, such as biomedical text.

We have also been working on constructing the infrastructure needed to support the production of data resources through the collection of web pages. We also allow the creation of document sets by selecting items from the publically available Common Crawl dataset[2].

On the **business development** side, the project partners have been following closely the developments and competitive landscape in the markets for text analytics, data-as-a-service, and software-as-a-service. Based on this research we will aim to come up with a pricing structure that is both competitive and sustainable.

User engagement is essential to the success of the AnnoMarket platform. Our technical design work has been closely informed by requirements collected through interviews with

[2] http://commoncrawl.org

many stakeholders, including both potential service providers and end users, coming from different industries. This should ensure the high relevance of the platform within then target communities, from the first moment it is open to the public. Following the release of the first platform prototype, the user engagement activity will be significantly ramped up with participation in events and increased on-line and social media presence. This will start with a demonstration of the first platform version during the 2013 conference of the Association for Computational Linguistics (http://acl2013.org/), the premier conference in its field, and a venue guaranteed to reach a very wide audience.

An early access version of the AnnoMarket platform will be made publicly available in the second half of 2013. The final AnnoMarket platform and associated services are due to be released to the public in mid-2014, potential users can register their interest on the project website at https://annomarket.eu.