This project is partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

D9.3 Exploitation Activity: Report Year 2

Licensing of the Resources Used for OpenBook Software Development

Author: Nikolay Pavlov

Affiliation: KDR

Date: 14 April 2014

Document Number: FIRST_D9.3_20140414

Status/Version: Approved – v1.2

Distribution Level: Public

Project Reference	*287607*
Project Acronym	*FIRST*
Project Full Title	*A Flexible Interactive Reading Support Tool*
Distribution Level	*Public*
Contractual Date of Delivery	*September 2013*
Actual Date of Delivery	*14 April 2014*
Document Number	*FIRST_D9.3_20140414*
Status & Version	*Final, Version 1.2*
Number of Pages	*41*
WP Contributing to the Deliverable	*WP9*
WP Task responsible	*Nikolay Pavlov*
Authors	*Nikolay Pavlov, Asen Rahnev, Dilyana Krushkova*
Other Contributors	*David Gil, Elena Lloret, Eduard Barbu, Nikki Sullings, Arlinda Cerga-Pashoja*
Reviewer	*Constantin Orasan*
EC Project Officer	*Magdalena Szwochertowska*
Keywords:	*Stakeholders, AT Market, NLP resources, Images, Licenses*

Abstract:

This document presents the main exploitation activities carried out during the second year of the FIRST project. The stakeholders' analysis and the different business models implementation from D9.2 are extended and described more precisely in this deliverable. A document which presents the resources used while developing the OpenBook software was also compiled. At the moment, 27 different resources are used, amongst which 23 NLP software products and 4 Image databases. They are licensed under 14 different types of licenses. The exploitation report presents general information about the licenses, some permissions and restrictions for the licensed products' exploitation and outlines conclusions for the research and commercial purposes of the FIRST project.

Contents

Introduction

OpenBook is the software product, developed as a result of the FIRST project whose aim is to open the doors to the world of online reading to people with Autism Spectrum Disorders (ASD). As explained in previous deliverables of the project, people with ASD may experience difficulties comprehending complex instructions, getting mislead by figurative language or the use of rare words or, simply, get distracted by secondary points touched upon in a document.

The previous version of the exploitation plan (D9.2) started considering the initial assessment of different business models and stakeholder analysis in order to provide information and guidelines that will help us to make some conclusions and decisions. The present document is the third and revised version of the exploitation plan developed after the EC reviewers' recommendations and will be followed by the final version at the end of the project (Month 36). D9.3 presents the exploitation activities carried out in the second year of the FIRST project.

The stakeholders' analysis and the different business models implementation from D9.2 are extended and described more precisely in this deliverable. A specific section about the potential geographic markets was added to D9.3. This section addresses where the Stakeholders of the OpenBook software are physically located. The main markets aimed are mainly within the EU, USA and Latin America, therefore, the magnitude of potential market (number of users with ASD/customers in these regions) is included here.

Information about the Assistive Technology Market and the OpenBook software as a part of it was added as a new section to the Exploitation plan after reviewers' recommendations. Stakeholders, assistive technology and different governmental programs for people with ASD are described in the terms of geographic markets.

The exploitation activities in the second year of the project also focused on identifying the various resources employed by the OpenBook software and determining the types of licenses under which they are available as each resource is presented separately. For the sake of completeness, each resource is briefly described, wherever possible, using the wording of the resource's authors and the main points of the license protecting it are presented. A discussion about how the resource can be used in the project is also provided.

Exploitation Scenarios

This section describes the possible methods of exploitation of the software product, result of the FIRST project, namely OpenBook. In general, there will be two possible different types of results that can be exploited in different ways:

1. Know-how and experience, gained through the project activities and research
2. Concrete software products and services

The first can be used for purposes like consultancy or other types of support for other similar projects. The OpenBook software and its separate components can be sold directly to different types of customers for various purposes (detailed description can be found in “Stakeholders analysis” section).

Several possible scenarios can be outlined for the use of the FIRST project software results.

Scenario 1

The first scenario assumes that the prototype developed within the scope of the FIRST project will be bundled as a whole concrete platform and will be exploited as a single product. That corresponds to the initial plans of the FIRST project and can be considered as the main purpose of the project. This scenario grants the following advantage: all ideas and functionality of the system are covered within one product and the end customers can gain from the whole experience.

Scenario 2

Due to the heterogeneity of the consortium it is likely that partners will exploit the parts they developed within the project by themselves. Research institutions can benefit from the exploitation of the components and services. Partners are experts in their own fields and might better address the needs of their customers/patients/students and markets when focusing on particular parts of the project development. The business models developed within the scope of the FIRST project will enable the IT partners to broaden their range of customers or strengthen their position within the market by having innovative advantages to their competitors. Research partners will be able to use innovations of the OpenBook software in their publications and studies which will facilitate to further research and development. Clinical partners will be provided with a tool that will improve and facilitate their work with the ASD patients and will lead to better understanding of autism as a whole.

Scenario 3

Combination of the parts provided can also be used as means of exploitation. Here, as combination is considered not only the use of the separate modules and services developed by the partners. Different software parts may be grouped together and larger parts of the concrete platform may be by combining the know-how and technology of the different partners. This scenario would allow the partners to provide a wider range of technologies and services and therefore address a wider range of customers.

Scenario 4

Using the OpenBook software as an abstract B2B platform – the software result of the efforts of the FIRST project consortium can be used as a starting point for other projects, related to NLP, text conversion or people with disabilities. The API of the software can be distributed without the frontend to serve as a base for other companies' projects. The abstract platform can be used by B2B clients for implementing parts with their own technology and in this sense it can be a driver for an evolving community around the ideas developed in the FIRST project. On the other side this also includes the exploitation of knowledge and experience gathered within the project duration like lessons learned and good practices.

Conclusion

Considering the different perspectives on how OpenBook tool and its components can be used we suggest that a mixture of the described scenarios is the right way for exploitation. Different scenarios will be more appropriate to different stakeholders. Further information about stakeholders and best scenario for each one of them can be found in section "Stakeholders analysis".

This section includes only a short plan of our initial exploitation plans. During the last year of the FIRST project development the partners will take a decision about the best exploitation scenario and then it will be described more precisely.

Freeware and Open Source

Some software components or the whole software product can be exploited as a freeware, so that the project achieves maximum impact: they should radiate as widely as possible so that the valuable lessons and experience gained by the FIRST project research groups can benefit others. Some OpenBook services can be distributed as an open source programs, ensuring that others can exploit the project's outputs. After initial discussion with the project partners it

became clear that none of the partners is against using the components created in relevant organization as open source.

The whole software product can also be exploited as a 'freemium'. Freemium (combination between the words "free" and "premium") has become a very popular business model in the software and internet services industry. This concept describes a business model where a core product is distributed for free and then generates revenue by selling premium products to a small percentage of free users. Adopting this new business model is a way of adapting to the changing market and the conditions of production.

The freemium version of the OpenBook software may be released under the following restrictions:

- "Lite", feature limited version of the software, not including all available features;
- Seat limited – can be used only on a single computer rather than across a whole network;
- Customer class limited – can be used freely only by a particular group of users (for example, only by families of people with ASD);
- Support limited – users of the "lite" version do not receive technical or other type of support;
- Time or bandwidth limited – it limits the time that free users can use the product. Two cases are available here:
    - At the end of the trial period some of the functionality stops working and user operates with the "lite" version of the software;
    - At the end of the trial period the whole product stops working.

Freemium depends on generating attention with the free product; then to sell premium products or services to some of the free users. In most cases only a small percent of the free users will purchase products. This is not a problem, as long as it is a small percentage of a large number. The advantages of this concept are obvious:

- the project results are distributed as wide as possible;
- our efforts would be beneficial for many users;
- there will be financial revenue from selling additional features/services.

Since additional distribution of a free product costs close to nothing, the success of a freemium business will increase with the number of people using it.

During this final year for the project the partners will take common decision whether there are appropriate software components to be distributed as a freeware and/or open source. The decision will be heavily influenced by the results of the tests that are to be conducted in the third year of the project.

Stakeholders Analysis

On D9.2 we started identifying some groups of stakeholders and described briefly each one of them. The following document presents a more detailed analysis of the stakeholders including information about how the results of the FIRST project can benefit each one of them.

Having different partners with different interests can also be seen as an advantage when focusing on more feasible and flexible scenarios.

Narrowing the field to key stakeholders and creating a profile for each one of them is a main objective of conducting a stakeholder analysis. The following groups can be affected by the project outcomes.

➢ The Commission

The results of projects funded through EU programs and initiatives need to achieve maximum impact: they should radiate as widely as possible so that the valuable lessons and experience gained by one group can benefit others. This involves preparing the ground for their work, carrying out their project while harvesting its results, distributing these results to the various end-users and stakeholders, and ensuring that others can exploit the project's output. Moreover, what is learnt from a project should inform future policy. The best option is that the project results reach and are used beyond the life of the project itself. That can be achieved if all exploitation scenarios are carried out successfully and also if parts of the software (or the whole software product) are distributed as freeware or an open source program (see the "Freeware and Open Source" section).

➢ The project partners

In Section 1 – General Principles of Annex II about the General Conditions from the FP7 Grant Agreement is mentioned that Foreground shall be the property of beneficiary carrying out the work generating that foreground. Where several beneficiaries have jointly carried out work generating foreground and where their respective share of the work cannot be ascertained, they shall have joint ownership of such foreground. They shall establish an agreement regarding the allocation and terms of exercising that joint ownership. Exploitation scenarios 2 and 3 are probably most appropriate. Each partner can exploit the components and services produced by the relevant organisation to improve and widen the scope of his work.

Collaboration with partners like Autism-Europe will be very useful in these terms in order to disseminate the information about the project results throughout already established

dissemination channels and thus, reach rest of the potential stakeholders in Europe and all over the world.

➢ Government institutions

Governmental organisations in most of the countries have several programs for improving the quality of life of people with disabilities including people with ASD. For example, European Commission calls for proposals in the area of employment, social affairs and inclusion present programs like VP/2010/017 "Pilot projects on employment of persons with autism spectrum disorders", aim to offer grant funding to applicants that may involve for-profit or not-for-profit organisations including public authorities, universities and research centers and civil society/stakeholder organisations. The presented program funds projects which develop or test approaches to improve access to and retention in the open labor market for persons with ASD in a practical manner. Other governmental support can be gained from some of the US programs for people with ASD. A total of 33 states and the District of Columbia have laws related to autism and insurance coverage. According to the Council for Affordable Health Insurance, an autism mandate increases the cost of health insurance by about 1 percent. However, if the incidence of autism continues to increase and as more services are covered, the cost of insurance may increase 1 to 3 percent. B2G approach can be used in this case.

The National Health Service in the UK is one of the world's largest publicly funded health services. The NHS is made of 129 Foundation Trusts, 58 Mental Health Trusts, 151 Primary Care Trusts that perform within the UK (England, Wales, Scotland and Ireland). The NHS employs more than 1.7m professionals who provide services to 63.2m people. Funding for the NHS comes directly from taxation and is granted to the Department of Health by Parliament. This budget for 2012/13 was around £108.9 billion.

The ideas, implemented by the OpenBook software can be also adopted by Community Associations that specialise in dealing with migration/migrant issues, Legal Advice organisations that work closely with migrants and the Citizen Advice Bureau (CAB) in UK (this is a very large Citizen Advice Bureau network with a CAB in every major UK town and city specialising in legal advice for adults). Scenario 1 is considered as the best one here.

➢ Private healthcare institutions

This group includes medical organisations, care and nursing homes and other centers specialized in Autism and other disorders that can affect reading comprehension. Some concrete institutions that belong to this group can be mentioned here:

- AyudaTec helps people with disabilities in Latin America by spreading the word about assistive technologies. AyudaTec products and services help people to communicate, learn, and live to their fullest potential. AyudaTec is supported by donations, companies

that contract for services, and through a voluntary Advisory Board. For more information about this stakeholder, visit the following link: http://www.ayudatec.org/.

- The Trust for the Americas is a non-profit organization affiliated with the Organization of American States (OAS). It was established in 1997 to promote public and private sector participation in social and economic development projects in Latin America and the Caribbean. Their initiatives, implemented through local partner organizations, seek to improve access to economic opportunities for vulnerable communities in the hemisphere. The Trust also promotes social inclusion and good governance.
  For more information about this stakeholder, visit the following link: http://trustfortheamericas.org/.
- Primary Care Trusts dealing with ASD in any shape or form;
- Strategic Health Authorities which are responsible for developing plans to improve health services in their local area.

These organisations create and carry out plans for ensuring that local health services are of a high quality and are performing well, for increasing the capacity of local health services so they can provide more services, and for making sure national priorities (for example, programs for improving autism services) are integrated into local health service plans.

As a result of the project, OpenBook will facilitate learning activities for children and adolescents with ASD and other reading disorders. It will also be in use of the specialists working with them as it reduces the time needed for preparation and adjustment of educational tasks and will increase student's autonomy. Carers will be able to simplify written information fast and convenient, compared to their previous experience. Scenarios 1 and 3 may be the most appropriate on this occasion.

➢ Research groups dealing with ASD

Some of the potential target research organisations are included in the list below:

- Research and Development Directorates across the UK (178 R&D departments in Primary Care Trusts, 293 R&D departments in Foundation Trusts);
- 25 Comprehensive Local Research Networks (CLRN's) across the UK;
- The 25 CLRN's combined form the Comprehensive Clinical Research Network (CCRN) governed by the National Institute of Health and Research (NIHR).
- CEDETI Centro de Desarrollo de Tecnologías de Inclusión – National University with 20% public support, ranked 3 in Latin America. Research areas covered by the University include medicine and social inclusion.
  http://www.cedeti.cl/.

These locally based Research Networks coordinate and facilitate/provide a wide range of support to the local research community. They have numerous specialities from Mental Health

Research, Cancer Research etc. The software produced can help scientists understand better people with ASD and to come up with new solutions in the general knowledge of ASD. The results and conclusions for the main obstacles in social inclusion of people with ASD point out possible directions for developing new helpful and accessible tools and programs for educational, therapeutic and communicational support to people with different types of learning, developmental or language disabilities. Scenario 1 is considered as the best option here.

➢ Software companies

Software companies can be reached by using B2B approach. The architecture of OpenBook enables us to easily create an Application Program Interface (API), based on SOAP web-services. These can be made available to businesses, which are interested in text simplification services. An alternative approach is to give businesses a premises based license, where they can install the API on their own servers, and extend it with their own resources for images and dictionaries as described in the 4th Scenario.

➢ Researchers in other reading assistive technology projects and research groups in NLP

The innovative language technologies services that will create exploitable results can be used by universities and other scientist groups or organisations to help in their research. The innovations developed within the scope of the FIRST project can be also useful for other similar projects sharing knowledge and experience, resources and innovations. Use of NLP components will be the best scenario in such cases.

➢ Users with ASD and their families

The conducted research presents information for the resources that people with ASD use to complete their everyday tasks and also for the challenges they meet in educational, interaction and other social situations. The facilitated reading will make the access to written information easier and enjoyable. This will widen the area of interests of the people with ASD and could be in support of communication with peers.

Cooperating with partners like Autism Europe can be considered a great advantage for the whole project, since it provides direct access to users with ASD. Users, their families and caregivers are involved at each stage of the software development to test and evaluate it. Their requirements will help the partners to improve and polish the interface and functionality of the system. Having involved users with ASD in the project from beginning to the end we can rely on their objective and useful feedback in order to fulfil their needs and requirements and thus carry out the project objectives effective and successfully.

First scenario is the one most appropriate option for this target group because they will need the whole functionality and experience while using the OpenBook software.

The poverty rate of people with a disability is 70% higher than average according to EC research (http://ec.europa.eu/news/justice/101115_en.htm). This is partly due to limited access to employment. Clearly related to access to employment is adequate access to education and training. Promoting accessibility is an important part of creating a culture of equal opportunities for all in the EU. But it also stands to benefit the economy as a whole. Boosting the industries that invest in accessible products and services will foster innovation and create jobs.

Therefore, it would be best if ASD users and their families are able to gain the software through government support or use it as a freemium. Our initial plans are to distribute the software freely to this target group.

➢ Distributors and intermediaries

The appropriate methods for distribution and advertising can easily help the software to gain enough popularity and to inform every potential customer about the product and its advantages. We can focus our attention on already established distribution channels and distributor companies - intermediaries. The collaboration with distributors and intermediaries can bring additional revenue to the manufacturer, due to their professional distributional methods, which contributes to increase of sales. Distributors can facilitate all aspects of the software exploitation and commercialization but will be most appropriate to distribute the whole system (Scenario 1).

➢ Schools and universities

Such organisations are considered as a secondary market for the software. The new technology can be brought to schools by the so called “resource teachers”. The resource teacher is a certified teacher who has a special qualification to teach students with learning disabilities (some of them with ASD). Such students require some modified education or more help in certain areas, most often in reading comprehension, but also in mathematics and other subjects. Such students are helped by resource teachers in order to keep up with the assigned workload. The OpenBook software will be very helpful for both teachers and students with special educational needs as it will improve their collaboration and facilitate reading comprehension, thus, improving the student’s performance in class. The new technology can be used to facilitate education of children in the early grades and university students with disabilities. The whole software product will be used in these cases (Scenario 1);

➢ Editors of books for children

Editors of children books can be also considered as a target group. The software will help them edit books in order to make them more accessible and easy to understand for children with special educational needs.

Assistive Technology Market

Overview

Assistive technology products are designed to assist people who, because of specific disabilities, would otherwise be unable to participate meaningfully in economic, social, political, cultural and other forms of activity in their communities according to BBC research about the Market for Disabled and Elderly Assistive Technologies. Assistive technology encompasses a broad range of devices, from "low-tech" products such as eyeglasses and large-print books, to technologically sophisticated products such as voice synthesizers, Braille readers and wireless monitoring devices. In ICT perspective assistive technology can be considered as hardware and software especially designed for disabled people. In general, the AT and specifically assistive ICT in the field of ICT, make these adjustments possible, such that people with a disability can enjoy the same access, use and ultimately the same rights as all citizens.

In the 21st century, software is a critical market for just about everyone, and disabled people are no exception. Rather, in many cases, these groups of end-users may be even more dependent on software developments and other technological developments as new tools to improve their accessibility to the world.

There has been substantial growth in the number of persons with disabilities, ASD including. As a result, this group is living longer independent or semi-independent lives with the help of assistive technologies.

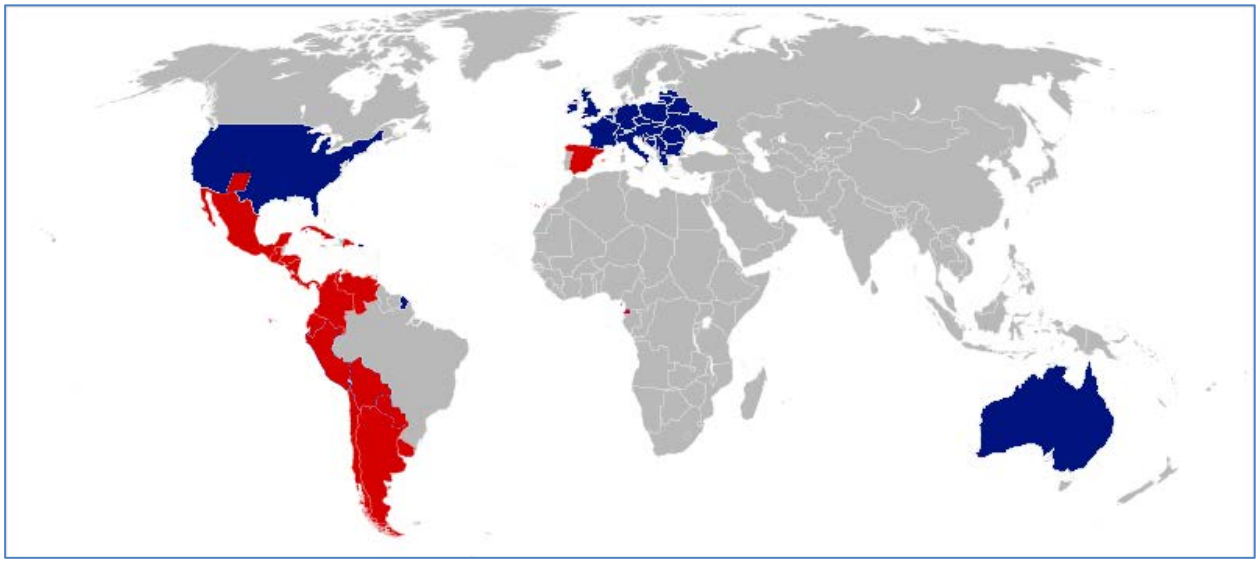
One of the problems for people with disabilities is that while IT is fast developing sector, new technology tends to throw up fresh barriers. Usually, technology companies neglect accessibility when developing new products and have to adapt them later for disabled people, which causes lot of efforts that could have been avoided if the adaptations had been part of the original specifications.

According to recent research about improving the software usability with assistive technologies (http://www.computerweekly.com/feature/Improve-usability-with-assistive-technology),

accessible IT does not need to be very complicated. Often it is just a matter of allowing a user to make adjustments to the size of type on screen or the color combinations in a display.

Assistive technology market in EU, USA and Spanish speaking countries

This section describes the ASD rates in the United States and the European Union countries considered as the main market of the project results. That's why the assistive technology market will be examined mainly in these regions. Considering that the OpenBook Software has a Spanish version, too, potential stakeholders can be found in Spanish speaking countries outlined here as a secondary market.



Autism Spectrum Disorder (ASD) is a significant public health challenge. Current estimates are that approximately 1% of the US and UK population has ASD, which, if one were to extrapolate, means five million people in EU member state countries are on the autism spectrum. The statistic is similar for the Spanish speaking countries, Australia and RSA – 1% of the population has some form of autism.

These current estimates show that ASD is more common than childhood cancer, juvenile diabetes and pediatric AIDS combined according to Autism speaks research. Awareness and prevalence of ASD has increased rapidly over the last ten years but there is still a lack of knowledge of the landscape of ASD in European countries. Many families and individuals on the autism spectrum have a daily struggle, so understanding, support and guidance is needed.

Progress in medical science as well as technology and healthcare, combined with demographic trends, societal evolution, and changing attitudes are driving the market for assistive technologies. If eyeglasses and contacts are excluded, the U.S. assistive technologies market was worth nearly $12.5 billion in 2012 and is expected to reach $13.2 billion in 2013 and $16.7 billion in 2018 for a CAGR of 4.8% between 2013 and 2018. These figures prove that Assistive

technology market is fast developing market niche as the number of people who will need such technologies is rising too.

Legislation

In general, the rights of people with disabilities feature clearly on the political agenda.

Most of the developed countries have special legislations that call the employers to make their services available to disabled people, provided it can be done at a reasonable cost. The United Nations Convention on the Rights of People with Disabilities and its optional protocol form an important legal framework for the achievement of equal rights and treatment of such people. The Convention puts forward the important concept of reasonable accommodation particularly for the work environment that focuses on ensuring an enabling environment for people with disabilities. The rights of disabled people to participate fully in all aspects of life, including the aim to "promote access for persons with disabilities to new information and communications technologies and systems, including the Internet" and to "promote the design, development, production and distribution of accessible information and communications technologies and systems at an early stage, so that these technologies and systems become accessible at minimum cost" are also covered. Similar legislation is the Disability Discrimination Act (UK) which made it unlawful to discriminate against people in respect of their disabilities in relation to employment, the provision of goods and services, education and transport.

The EU has been active for many years in the field of disability, covering different policy fields such as information society, education, employment, health, etc. Recently the European Disability Strategy 2010-2020 was adopted. Particularly, the new disability strategy stresses the importance of "improving the availability and choice of assistive technologies".

The societal shift that led to persons with disabilities and the elderly joining the mainstream of daily life in America was solidified in 1990 with the passage of the Americans with Disabilities Act. The act forced all the major institutions of American life, including workplaces, government and commercial buildings, educational institutions, and public transportation to accommodate the needs of the elderly and persons with disabilities. The next big shift for this marketplace will be seen after the passage and implementation of the Patient Protection and Affordable Care Act of 2010.

That's why many IT companies are under increasing pressure to ensure that their systems can be used by everyone.

Conclusions

Most EU Member States have support schemes in place helping the disabled people to integrate socially or to find their right place on the labour market. The presence of these support schemes has a profound impact on the Internal Market for assistive ICT.

Recent researches show that the AT market is very rapidly developing but is still in the making. Disabled people are the ultimate consumers of assistive ICT, but often are not free to choose or adapt the assistive ICT product that best suits their individual needs. At present, about 80% of the software that is available for AT applications is available only in English. Language is also an important issue because it serves as an important barrier between the different country markets of the EU AT ICT industry. Considered as a part of this market, the OpenBook software can significantly contribute to the needs of this market allowing people with ASD to participate in many social activities. Being available in 3 languages at the start, software can remove additional barriers to the public interaction of users with ASD. It will also have an impact on companies and intermediaries that would focus directly on their consumers, people with disabilities and other public service providers.

Types of licenses

Different software may be licensed under a variety of licensing schemes. For the users of software there are various licenses ranging from very restrictive proprietary licenses to open-source licenses. There are also different licensing schemes for access to and use of source code. To address special intellectual property issues regarding source code, open-source licenses, and special licenses, such as copyleft, have been created.

Not all software is licensed, or even copyrighted. It may be published without an accompanying license (such as License-Free Software) in which case it remains copyrighted, its distribution is subject to ordinary copyright law, and its sale is subject to ordinary sales law. In some countries, software may also be released to the public domain. The public domain consists of works that are available for public use. In this case it is not copyrighted and the notion of a copyright license simply does not apply at all although the other parts of a software license, including warranty provisions, will still apply to the sale of such software.

Salable graphic libraries can be found through the Internet. Stock photos which are ready-made images licensable for use after paying a fee can also be considered as appropriate for the project. However, the authors/licenses of many images within the web space remain unknown.

System Software

The prototype of OpenBook was developed as a web-based application. The following resources were used for the development of the system:

- System software:

  - Operating systems –Windows Server 2008 or newer, Linux;
  - Relational database management systems – Microsoft SQL Server 2008;
  - Software runtime systems - C#, ASP.NET, .NET Framework 4.0;
- Software libraries – Java Script, jQuery, WCF;
- Language resources – dictionaries, frequency lists, Wikipedia, etc.;
- Graphic resources – image databases, Google, Bing.

More information about the different components of the system can be found below.

OpenBook consists of three primary software components:

1. NLP services, running under Linux operating system.
2. Integration Engine, running under Windows Server 2008 (or later) operating system.
3. Front-end web application, running under Windows Server 2008 (or later) operating system.

Operating Systems

Linux is an operating system under the model of free and open-source software. This means that it can be obtained, installed and used without paying license costs. Therefore, it implies no risk for either the research or the commercial continuity of the FIRST project.

Windows Server is a commercial operating system, developed by the Microsoft Corporation. The current version of Windows Server as per the end of the second year of the FIRST project is 2012. Therefore, only this version will be considered in the deliverable. The licensing scheme of Windows Server is complex, and is out of the scope of the current study. This document will summarise only the requirements for the FIRST project.

For the needs of the FIRST project, both research and initial commercial activities can be successfully executed using the Essentials edition of Windows Server. This edition is suitable for running line-of-business applications and does not require additional Client Access Licenses (CALs). The price of the Essentials edition of Windows Server 2012 set by Microsoft Corporation is $501.

We consider the price as being a negligible risk to the commercial exploitation of the FIRST project.

Database

In addition, the Integration Engine requires a relational database instrument to operate. The implementation relies on Microsoft SQL Server, version 2008 or newer. The technical requirements of OpenBook define that the Microsoft SQL Server Express Edition is sufficient for both research and initial commercial exploitation of the FIRST project.

Microsoft SQL Server Express Edition is a closed-source, but free database. It includes 10GB of storage per database. More information about Microsoft SQL Server Express Edition can be found here: http://www.microsoft.com/en-us/sqlserver/editions/2012-editions/express.aspx.

We consider that Microsoft SQL Server Express Edition poses no risk to either the research or commercial exploitation activities.

OpenBook will store user accounts, configuration preferences and documents, but will not store images. Considering that we expect the average document to be up to 50K in size, we can safely expect this edition of Microsoft SQL Server to handle more than 200,000 documents, which is largely sufficient for the initial stages of commercial exploitation.

In future, if data storage demand grows, but the price of the SQL Server is still too high, the database can be replaced with an open-source relational storage such as MySQL, or Firebird.

Another storage alternative would be to use the cloud version of Microsoft SQL Server, where it is possible to pay only for the storage required.

Licenses of used NLP software applications

1. Creative Commons Attribution 3.0 Unported License

Creative Commons (CC) is a non-profit organisation devoted to expanding the range of creative works available for others to build upon legally and to share. The organisation has released several copyright licenses, known as Creative Commons licenses, free of charge to the public. These licenses allow creators to communicate which rights they reserve, and which rights they waive for the benefit of recipients or other creators.

Creative Commons Attribution 3.0 Unported License allows us:

- to share — to copy, distribute and transmit the work;
- to remix — to adapt the work;
- to make commercial use of the work.

An easy-to-understand one-page explanation of rights, with associated visual symbols, explains the specifics of each Creative Commons license. Summary of the Creative Commons Attribution 3.0 Unported License can be found here:

http://creativecommons.org/licenses/by/3.0/

The full text of the license can be found here:

http://creativecommons.org/licenses/by/3.0/legalcode

Licensed software products: The Paraphase Database, MultiWordNet, Wiktionary, WordNet Domains.

Conclusion:

Using software products licensed under the Creative Commons Attribution 3.0 Unported License will be useful for the development of NLP web services and will not cause problematic issues for the Project and for the commercial exploitation of the OpenBook software.

2. Creative Commons Attribution-NonCommercial-Share Alike 3.0 License

This license is another one from the Creative Commons group.

Creative Commons Attribution-NonCommercial-Share Alike 3.0 License allows us:

- to share - to copy, distribute and transmit the work;
- to remix - to adapt the work.

The above rights may be exercised in all media and formats.

Restrictions:

We can distribute or publicly perform our work only under the terms of this License. We must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work distributed or publicly performed – i.e. the new work should be made available under the same license terms.

This License is only for Noncommercial use — We may not use this work for commercial purposes. ("You may not exercise any of the rights granted in the previous section in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works…" – extraction from the license full text.)

The full text of the license can be found here:

http://creativecommons.org/licenses/by-nc-sa/3.0/legalcode

Notice — For any reuse or distribution, we must make clear to others the license terms of this work. The best way to do this is with a link to the summarised version of the license - http://creativecommons.org/licenses/by-nc-sa/3.0/.

Licensed software products: Babelnet, Bulgarian National Corpus (BulNC), Wikipedia Images.

Conclusion:

Using software products licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License will not cause problems as long as FIRST is a research project but some copyright issues may arise for the commercial use of OpenBook. We will explore the options to legally acquire the right to use this software for commercial purposes.

3. MS-NC-NoReD (META-SHARE Non Commercial No Redistribution) License Agreement

This license grants a worldwide, non-exclusive, Non Commercial license, clear of any third parties rights, to use the Resource under the terms and conditions defined.

It allows us to:

- copy the Resource, create Derivatives or incorporate the Resource or the Derivatives into a Collective Work;
- extract and re-utilise the whole or substantial parts of the Resource; copy Derivatives or the Resource as incorporated in any Collective Work;
- distribute any Derivative product or service based on all or a substantial part of the Resource for Non Commercial purposes.

To use software products licensed under this agreement for research purposes we should sign the following agreement: http://www.bultreebank.org/dpbtb/dp-btb-license.htm

Some references should be cited in any publication reporting about results obtained on these data.

However, this license has many restrictions. We are not allowed to:

- use the Resource in any other way other than as necessary or desirable for the purposes of our own internal Non Commercial language engineering research activities or our own internal Non Commercial technology development;
- make available to the public all or any substantial part of the contents of the Resource, by the distribution of copies, by renting, leasing or any other form of distribution;
- sublicense the Resource; or
- use the Resource or its Derivatives for commercial purposes.

Here is the full text of the license:

http://metashare.nb.no/repository/download/a8d0978e395711e2b66e001708556d5aaf37775ab9104bc79e554d520eeefda1/

For more information on how to obtain the selected resource under these license terms the resource maintainer should be contacted.

Licensed software products: Bulgarian Treebank word frequency list.

Conclusion:

The software licensed under MS-NC-NoReD can be used for research purposes only after signing an agreement. The data or part of the data in its original or modified form cannot be distributed. The commercial use and the redistribution of such software products is not allowed.

4. GNU General Public License

The GNU General Public License is a free, copyleft license. Copyleft is a general method for making a program (or other product) free, and requires all modified and extended versions of the program to be free as well. Here, 'free' does not necessarily mean free of charge, but free as in freely available to be modified.

The GNU General Public License is intended to guarantee the users' freedom to share and change all versions of a program - to make sure it remains free software for all its users. The General Public Licenses guarantee freedom to distribute copies of free software (and charge for them if necessary). Thee recipients of such software products should gain the same freedoms as the user of the code. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

The full text of the license can be found here: http://www.gnu.org/licenses/gpl.html

Licensed software products: Freeling 3.0, Weka, BalkaNet.

Conclusion:

Using software licensed under GNU General Public License will not cause problems in terms of commercial use of the OpenBook software, but may not be appropriate because of its "copyleft" nature. The problem here is that we may need to make the whole source code of the software available to the mass audience. Software, licensed under GNU GPL will not cause problematic issues for the research purposes of FIRST.

5. GNU Lesser General Public License

This version of the GNU Lesser General Public License incorporates the terms and conditions of version 3 of the GNU General Public License, supplemented by some additional permissions.

The LGPL allows developers and companies to use and integrate LGPL software into their own (even proprietary) software without being required by the terms of a strong copyleft to release the source code of their own software-parts. Merely the LGPL software-parts need to be modifiable by end-users (via source code availability): therefore, in the case of proprietary

software, the LGPL-parts are usually used in the form of a shared library (e.g. DLL), so that there is a clear separation between the proprietary parts and open source LGPL parts.

The "Lesser" in the title of the license is used, to show that LGPL cannot guarantee end users complete freedom in the use of software, because only the LGPL-parts (but not any proprietary software-parts) guarantee end users the access to source code and therefore the freedom of modification.

The full text of the license can be found here: http://www.gnu.org/licenses/lgpl.html

Licensed software products: Gate 7.0, JWPL.

Conclusion:

Using software licensed under GNU Lesser General Public License will not cause problematic issues for commercial purposes if the (modified) source code of the used parts of code is accessible for the mass audience. This software can be used freely for research purposes.

6. Common Public License

Subject to the terms of this Agreement, each Contributor hereby grants Recipient a non-exclusive, worldwide, royalty-free patent license under Licensed Patents to make, use, sell, offer to sell, import and otherwise transfer the Contribution of such Contributor, if any, in source code and object code form. This patent license shall apply to the combination of the Contribution and the Program if, at the time the Contribution is added by the Contributor, such addition of the Contribution causes such combination to be covered by the Licensed Patents. The patent license shall not apply to any other combinations which include the Contribution. No hardware per se is licensed hereunder.

The full text of the license can be found here: http://opensource.org/licenses/cpl1.0.php

Licensed software products: MALLET.

Conclusion:

Using software licensed under Common Public License will not cause problematic issues for the project development and for the commercial use of the OpenBook software.

7. BSD 3-Clause License

BSD licenses are a family of permissive free software licenses imposing minimal restrictions on the redistribution of covered software. The only restrictions placed on users of software released under a typical BSD license are that if they redistribute such software in any form, with or without modification, they must include in the redistribution the original copyright notice, a list of two simple restrictions and a disclaimer of liability.

This version allows unlimited redistribution for any purpose as long as its copyright notices and the license's disclaimers of warranty are maintained. The license also contains a clause restricting use of the names of contributors for endorsement of a derived work without specific permission.

The full text of the license can be found here: http://opensource.org/licenses/BSD-3-Clause

Licensed software products: JWNL (Java WordNet Library) 1.4.1.

Conclusion:

The software licensed under BSD license will not cause problematic issues for the project development. It can also be used for commercial purposes in derivative works if we retain the original copyright notice.

8. Apache License, Version 2.0

Apache License, Version 2.0 – Subject to the terms and conditions of this License, each Contributor hereby grants a perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare Derivative Works of, publicly display, publicly perform, sublicense, and distribute the Work and such Derivative Works in Source or Object form.

We are allowed to reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that we meet the following conditions:

- to provide any other recipients of the Work or Derivative Works with a copy of this License; and
- to cause any modified files to carry prominent notices stating that there are files changed; and
- to retain, in the Source form of any Derivative Works that we distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, etc.

Here can be found the full text of the license with some additional information about how to apply Apache License to our work: http://www.apache.org/licenses/LICENSE-2.0

Licensed software products: Apache Lucene TM.

Conclusion:

Software licensed under Apache License can be used for commercial, research and open-source programs provided that we meet the conditions mentioned above.

9. GNU Free Documentation License

The GNU Free Documentation License (GNU FDL or simply GFDL) is a copyleft license for free documentation. It is similar to the GNU General Public License, giving readers the rights to copy, redistribute, and modify a work and requires all copies and derivatives to be available under the same license. Copies may also be sold commercially, but, if produced in larger quantities (greater than 100), the original document or source code must be made available to the work's recipient.

The full text of the license can be found here: http://www.gnu.org/copyleft/fdl.html

Licensed software products: Wiktionary.

Conclusion:

Using software licensed under GNU General Public License will not cause problems in terms of commercial use of the OpenBook software, but may not be appropriate because of its “copyleft” nature. Such software will not cause problematic issues for the research purposes of FIRST.

10. Software licensed under the Linguistic Data Consortium License

The Linguistic Data Consortium supports language-related education, research and technology development by creating and sharing linguistic resources: data, tools and standards. It provides two types of licensing agreements – license for members of the consortium and non-member license.

LDC Membership Agreements

For Members, use of most data is governed by the membership agreement. There are distinct license agreements for the four categories of LDC membership: "for-profit" (corporate or commercial), "not-for-profit" (academic or non-commercial), U.S. Government, and LDC Online (not-for-profit and government entities only).

Here are the agreements to be filled and signed in two copies:

- For Not-For-Profit organisations;
- For For-Profit organisations;
- For U.S. Government Entities;
- For LDC Online Membership (not-for-profit and government entities).

Here is the list of the corpora available that require user licenses, along with links to the license forms: http://www.ldc.upenn.edu/Membership/Agreements/licenses/

For Non-members, use of most data is governed by the LDC User Agreement for Non-members. This user license agreement permits non-commercial linguistic education and research use of data.

Here is the agreement which should be filled and signed for the non-members license - http://www.ldc.upenn.edu/Membership/Agreements/licenses/genericlicense.pdf

Licensed software: NLP tasks from SemEval and Senseval evaluations: Coreference Resolution in Multiple Languages, English and Spanish Word Sense Disambiguation, Multilingual Word sense disambiguation task.

Conclusion:

Semeval and Senseval tasks can be used for non-commercial linguistic education and research purposes, but not for commercial.

11. Software licensed by Universitaet Stuttgart

The Institut fuer maschinelle Sprachverarbeitung, Universitaet Stuttgart, develops tools for text annotation.

The software used for development of the OpenBook software is called TreeTagger and is licensed under the terms of the Institut fuer maschinelle Sprachverarbeitung agreement.

The TreeTagger software can be used for evaluation, research and teaching purposes. Any other usage of the system (in particular for commercial purposes) is forbidden.

By downloading the software, user agrees to the terms stated there: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/Tagger-Licence

Licensed software: TreeTagger.

Conclusion:

This software is freely available for research, education and evaluation but any other usage of the system (in particular for commercial purposes) is forbidden.

12. SIL Open Font License 1.1

The SIL Open Font License (or OFL in short) is a free and open-source license designed for fonts by SIL International for use with some of their Unicode fonts. The Open Font License is a free software license, and as such permits the fonts to be used, modified, and distributed freely (so long as the resulting fonts remain under the Open Font License). The License permits covered fonts to be freely embedded in documents under any terms, but it requires that fonts be packaged with software if they are sold.

The full text of the license may be found here: http://opensource.org/licenses/OFL-1.1

Licensed software: Font Awesome.

Conclusion:

There is no problem for us to use the fonts licensed under SIL OFL neither for research purposes nor for the commercial version of OpenBook.

13. ImageNet Licensing Agreement

ImageNet is an image database organized according to the WordNet hierarchy.

ImageNet does not own the copyright of the images. ImageNet only provides thumbnails and URLs of images, in a way similar to what image search engines do. The images in their original resolutions may be subject to copyright, so they cannot be made publicly available on the server. Researchers/educators who wish to have a copy of the original images for non-commercial research and/or educational use can be provided access through the ImageNet site, under certain conditions and at their discretion. First, the researcher should send a request to the ImageNet support, then he should be waiting for approval. After the request is approved, the researcher should agree to sign the following terms of access: http://www.image-net.org/download-faq

Licensed software: ImageNet.

Conclusion:

The Database can be used only for non-commercial research and educational purposes. The commercial use of the images will be a problem, because each image may be a subject to different copyright.

14. Oxford Text Archive License

By using The Oxford Text Archive we are bound by the following Terms & Conditions:

“All material supplied via the University of Oxford Text Archive (OTA) is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.”

The Oxford Text Archive will use your data in accordance with the Data Protection Act, as set out in its Privacy Statement.

Additional information about Terms and Conditions can be found here:

http://ota.ahds.ac.uk/scripts/download.php?otaid=0668

Licensed software: Kucera and Francis frequency list.

Conclusion:

Software licensed under Oxford Text Archive is freely available for non-commercial use provided that a specific header is included in its entirety with any copy distributed. Duplication or sale of all or part of any of the Data Collections is not permitted.

Used NLP software applications

1. The Paraphrase Database

The paraphrase database (PPDB) of The Center for Language and Speech Processing from Johns Hopkins University can be used in the syntactic simplification task in the First project (1.0 release will be soon available for English and Spanish). Its English portion, PPDB: Eng, contains over 220 million paraphrase pairs, consisting of 73 million phrasal and 8 million lexical paraphrases, as well as 140 million paraphrase patterns, which capture many meaning-preserving syntactic transformations. The database for Spanish contains 196 million Spanish paraphrases.

Source for more information: http://aclweb.org/anthology//N/N13/N13-1092.pdf

PPDB is licensed under a **Creative Commons Attribution 3.0 Unported License** which allows us to use it for both research and commercial purposes.

2. Babelnet

BabelNet is a multilingual lexicalised semantic network and ontology. BabelNet was automatically created by linking Wikipedia to WordNet. The integration is performed by an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of statistical machine translation. The result is an "encyclopedic dictionary" that provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations.

Source for more information: http://en.wikipedia.org/wiki/BabelNet

BabelNet and its API are licensed under a **Creative Commons Attribution-Noncommercial-Share Alike 3.0 License** which allows only non-commercial use**.**

3. Bulgarian National Corpus

Bulgarian National Corpus (BulNC) is a large publicly available corpus designed as a uniform framework for texts of different modality (written and spoken), period, and number of languages (monolingual and parallel). It is constantly enlarged and developed. The corpus design requires a clear-cut structure based on an explicit description of sample categories and explicit mapping between parallel samples in different languages.

Source for more information: http://metashare.ibl.bas.bg/repository/browse/bulgarian-national-corpus/817c127064aa11e281b65cf3fcb88b705d83aefb9d21409dbaf029c8dddcce00/

BulNC is licensed under **Creative Commons Attribution-Noncommercial-Share Alike 3.0 License** which allows only non-commercial use.

4. Bulgarian Treebank word frequency list

BulTreeBank project aims to develop a high quality set of syntactic trees for Bulgarian within the framework of Head-driven Phrase Structure Grammar (HPSG). Bulgarian Treebank is a text corpus in which each sentence has been annotated with syntactic structure (represented as a tree).

BulTreeBank Project was funded by the Volkswagen Stiftung, Federal Republic of Germany under the Programme "Cooperation with Natural and Engineering Scientists in Central and Eastern Europe".

Source for more information: http://www.bultreebank.org/

The Dependency and the Morphologically Annotated Parts of BulTreeBank are available under **MS-NC-NoReD (META-SHARE NonCommercial NoRedistribution) License Agreement** and can be used for research purposes only after signing an agreement.

5. Freeling 3.0

FreeLing is a system for the linguistic analysis of text developed at the Universitat Politècnica de Catalunya. The latest version of the system contains extensive language data for Spanish, Catalan, Galician, English, Italian, Welsh, Portuguese, Russian and Ancient Spanish. The system takes the form of a library, which can be called from within a computer program. There is also an application program available, called analyzer.exe, which allows most of the functionality of FreeLing to be used.

Source for more information: http://nlp.lsi.upc.edu/publications/papers/padro12.pdf

FreeLing is released under the **GNU General Public License** of the Free Software Foundation which will not cause problematic issues for the research purposes of FIRST but may not be appropriate because of its “copyleft” nature.

6. Gate 7.0

GATE - (General Architecture for Text Engineering) is the world's most popular software platform for language engineering, developed by the NLP group of the University of Sheffield.

The various scientific and engineering disciplines to which GATE is relevant are:

- Computational Linguistics: part of the science of language that uses computation as an investigative tool;
- Natural Language Processing: part of the science of computation whose subject matter is data structures and algorithms for human language processing;
- Language Engineering: building language processing systems whose cost and outputs are measurable and predictable.

The Flexible Gazetteer consists of a set of lists containing names of entities such as cities, organisations, days of the week, etc. These lists are used to find occurrences of these names in text, e.g. for the task of named entity recognition. The word 'gazetteer' is often used interchangeably for both the set of entity lists and for the processing resource that makes use of those lists to find occurrences of the names in text.

Source for more information: http://gate.ac.uk/download/

Gate 7.0 is licensed under **GNU Lesser General Public License** which will not cause problematic issues for commercial purposes if the (modified) source code of the used parts of code is accessible to the mass audience. This software can be used freely for research purposes.

7. MALLET

MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, etc.

MALLET includes sophisticated tools for document classification: efficient routines for converting text to "features", a wide variety of algorithms and code for evaluating classifier performance using several commonly used metrics.

Source for more information: http://mallet.cs.umass.edu/

MALLET is licensed under **Common Public License.** We are welcome to use the code under the terms of the license for research or commercial purposes, however we should acknowledge its use with a citation: "McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu. 2002."

8. Diccionario de la Real Academia de la Lengua Espanola

The Diccionario de la lengua española de la Real Academia Española or DRAE is the most authoritative dictionary of the Spanish language. It is produced, edited, and published by the

Real Academia Española (RAE – Royal Spanish Academy). The Real Academia Española (English: Royal Spanish Academy), generally abbreviated as RAE, is the official royal institution responsible for regulating the Spanish language.

The only task used from Real Academia is the list of frequent words that is available through their webpage and can be freely downloaded (http://corpus.rae.es/lfrecuencias.html).

The dictionary itself is not integrated in any web services at the moment. It cannot be freely downloaded. DRAE can be purchased for nearly $ 250.

http://www.amazon.com/Diccionario-Lengua-Espanola-Academia-CD-ROM/dp/0828848874

9. JWNL (Java WordNet Library) 1.4.1

JWNL is an API for accessing WordNet-style relational dictionaries. It also provides functionality such as relationship discovery and morphological processing. WordNet is a semantic lexicon for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

Source for more information:
http://www.stanford.edu/class/archive/cs/cs276a/cs276a.1032/projects/docs/jwnl/javadoc/

JWNL is licensed under **BSD 3-Clause License** which will not cause problematic issues for the project development. It can also be used for commercial purposes in derivative works if we retain the original copyright notice.

10. JWPL

JWPL (Java Wikipedia Library) is a free, Java-based application programming interface that allows accessing all information in Wikipedia.

Source for more information: https://code.google.com/p/jwpl/

JWPL is licensed under **GNU Lesser General Public License,** which will not cause problematic issues for commercial purposes if the (modified) source code of the used parts of code is accessible to the mass audience. This software can be used freely for research purposes.

11. Kucera-Francis word frequency list

Word lists by frequency are lists of a language's words grouped by frequency of occurrence within some given text corpus. A word list by frequency "provides a rational basis for making sure that learners get the best return for their vocabulary learning effort", (Nation 1997) but is mainly intended for course writers, not directly for learners.

The Brown (**Francis and Kucera**, 1982) LOB and related corpora now contain 1,000,000 words from a written corpora representing different dialects of English. These sources are used to produce frequency lists.

Kucera and Francis frequency list is freely available for non-commercial use provided that this header is included in its entirety with any copy distributed: "Kuçera-Francis wordlist [Electronic resource] : [a] frequency count of the Brown corpus of present day American English".

Duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated for research use or educational purposes in electronic or print form.

Kucera-Francis norms can also be found as part of the package 'WordPools', which collects several classical word pools used most often to provide lists of words in psychological studies of learning and memory. This package is available under **GPL-2 License**.

http://cran.r-project.org/web/packages/WordPools/WordPools.pdf

12. Lucene

Apache LuceneTM is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

Source for more information: http://lucene.apache.org/core/

Apache Lucene is available as open-source software under the **Apache License** which lets you use Lucene in both commercial and open-source programs.

13. MultiWordNet

MultiWordNet is a multilingual lexical database which includes information about English and Italian words. The database contains information about the following aspects of the English and Italian lexica:

- lexical relations between words;
- semantic relations between lexical concepts (synsets);
- correspondences between Italian and English lexical concepts;
- semantic fields (domains).

The basic lexical relationship in MultiWordNet is lexical synonymy. Groups of synonyms are used to identify lexical concepts, which are called synsets.

Source for more information: http://catalog.elra.info/product_info.php?products_id=644

MultiWordNet is licensed under a **Creative Commons Attribution 3.0 Unported License** which allows us to copy, distribute, transmit, adapt and to make commercial use of the work.

14. Newswire Corpus

The Newswire Corpus is used for extraction of texts relevant to different domains.

The Newswire Corpus is only used for testing services and improving them. University de Alicante collected this corpus for being able to have a collection of documents to test the developed modules.

The Newswire Corpus can be freely used for research purposes.

15. SemEval (task 1) and Senseval-3

SemEval (Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems. The evaluations are intended to explore the nature of meaning in language.

SemEval is focused on word sense disambiguation, each time growing in the number of languages offered in the tasks and in the number of participating teams.

Senseval-3 workshop took place in March-April 2004, followed by a workshop held in July 2004 in Barcelona, in conjunction with ACL 2004. Senseval-3 included 14 different tasks for core word sense disambiguation, as well as identification of semantic roles, multilingual annotations, logic forms, subcategorisation acquisition.

Most of the data sets used in this exercise are **available for download** (trial, train, and/or test data) and results of all participating systems are also available for download.

SemEval is free for members of the **Linguistic Data Consortium**. The different NLP tasks are licensed under different LDG licenses. More information can be found in the License section.

The datasets of the following tasks were used during the NLP services development:

- For Spanish coreference - SemEval-2010 corpus for Task 1: Coreference Resolution in Multiple Languages - http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2011T01
  Here can be used the following licenses: Subscription Members, Standard Members and Non-Members;
- For English and Spanish Word Sense Disambiguation - Senseval-3 (year 2004) - http://www.senseval.org/senseval3/tasks.html, freely available resources;
- Semeval 2013 for the Multilingual Word sense disambiguation task - http://www.cs.york.ac.uk/semeval-2013/task12/

  Test data is publicly available and free for download.

16. TreeTagger

The aim of the **TreeTagger** tool is to annotate text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger is successfully used to tag different languages, amongst which are German, English, French, Italian, Spanish, Bulgarian, Russian and many other texts. The tool is adaptable to other languages if a lexicon and a manually tagged training corpus are available.

Source for more information: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

The TreeTagger is licensed under **TreeTagger License** of the The Institut fuer maschinelle Sprachverarbeitung, Universitaet Stuttgart which allows use the system for research, education and evaluation but any other usage (in particular for commercial purposes) is forbidden.

17. Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java. The Weka workbench contains a collection of visualisation tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Source for more information: http://en.wikipedia.org/wiki/Weka_(machine_learning)

Weka is free software available under the **GNU General Public License** which will not cause problematic issues for the research purposes of FIRST but may not be appropriate because of its “copyleft” nature.

18. Wiktionary

Wiktionary is a collaborative project to produce a free-content multilingual dictionary which aims to describe all words of all languages using definitions in English.

Designed as the lexical companion to Wikipedia, Wiktionary has grown beyond a standard dictionary and now includes a thesaurus, a rhyme guide, phrase books, language statistics and extensive appendices.

Source for more information: http://en.wiktionary.org/wiki/Wiktionary:Main_Page

Original text of Wiktionary entries is dual-licensed to the public under both the **Creative Commons Attribution-ShareAlike 3.0 Unported License**, as well as **GNU Free Documentation License (GFDL)** and can be used both for research and commercial purposes.

19. WN Domains

WordNet Domains is a lexical resource created automatically by expanding WordNet with domain labels. WordNet Synsets have been annotated with at least one semantic domain label, selected from a set of about two hundred labels structured according the WordNet Domain Hierarchy.

WordNet Domains includes WordNet-Affect, an additional hierarchy of "affective" domain labels, with which the synsets representing affective concepts are further annotated.

Source for more information: http://wndomains.fbk.eu/

WordNet Domains is licensed under a **Creative Commons Attribution 3.0 Unported License** and is distributed, both for research and for commercial purposes.

20. BalkaNet

BalkaNet is a multilingual lexical database consisting of WordNets in several European languages. The database development is still in progress. New features will be implemented to ensure linking of conceptual equivalencies across WordNets to the development of an inter-networked WordNet Management so that each partner retains full responsibility and independence of his local WordNet. At the same time they will be able to view other WordNets and check their compatibility.

Source for more information: http://www.dblab.upatras.gr/balkanet/pubs/GWA_paper_03.pdf

BalkaNet is licensed under **GNU General Public License** which will not cause problematic issues for the research purposes of FIRST but may not be appropriate because of its “copyleft” nature.

21. Font Awesome

Font Awesome provides vector icons that can easily be customised — size, color, drop shadow, and anything that can be done with CSS. Some of the icons were used in the OpenBook interface for users.

Source for more information: http://fortawesome.github.io/Font-Awesome/

Font Awesome is fully open-source and is GPL compatible. It can be used for commercial projects, open-source projects, etc. The **SIL OFL 1.1 License** allows us to use the Font Awesome for research and commercial purposes if it is packed with a software.

22. WordNet

WordNet is a lexical database for the English language which groups English words into sets of synonyms called synsets, provides short, general definitions. It also records various semantic relations between the synsets. The purpose is to produce a combination of dictionary and

thesaurus that is more intuitively usable and to support automatic text analysis and artificial intelligence applications.

Source for more information: http://wordnet.princeton.edu/

The database and software tools have been released under a **BSD style license** and can be downloaded and used freely. The database can also be browsed online.

Permission to use, copy, modify and distribute this software and database and its documentation for any purpose and without fee or royalty is hereby granted.

23. Wikipedia

Wikipedia is a multilingual, web-based, free-content encyclopedia based on an openly editable model. The name "Wikipedia" is a combination of the words wiki (a technology for creating collaborative websites, from the Hawaiian word wiki, meaning "quick") and encyclopedia. Wikipedia's articles provide links designed to guide the user to related pages with additional information.

Wikipedia is written collaboratively by largely anonymous Internet volunteers who write without pay. Users can contribute anonymously, under a pseudonym, or, if they choose to, with their real identity.

Source for more information: http://en.wikipedia.org/wiki/Wikipedia:About

Most of Wikipedia's text and many of its images are dual-licensed under the **Creative Commons Attribution-Sharealike 3.0 Unported License (CC-BY-SA)** and the **GNU Free Documentation License (GFDL)** and can be used both for research and commercial purposes.

24. ImageNet

ImageNet is an image dataset organized according to the WordNet hierarchy. There are more than 100,000 synsets in WordNet, majority of them are nouns (80,000+). ImageNet we aims to provide on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated. In its completion ImageNet will offer tens of millions of cleanly sorted images for most of the concepts in the WordNet hierarchy.

The Database can be used only for non-commercial research and educational purposes.

Source for more information: http://www-cs.stanford.edu/content/imagenet-large-scale-hierarchical-image-database

25. Idiom Dictionaries for Spanish

Each site contains freely available resources for compiled lists of Spanish Idioms. Here are listed the resources used:

- Dictionary of Spanish Idioms and Expressions - http://www.spanish-learning-corner.com/spanish-idioms.html
- Language Realm (free translation resources) – Spanish Idioms - http://www.languagerealm.com/spanish/spanishidioms.php
- Spanish idioms with their English equivalents embracing nearly ten thousand phrases (1899) - http://archive.org/details/spanishidiomswit00beck
- Dichos populares - http://www.ciudad-real.es/varios/dichos/a.php
- Dichos y frases hechas - http://iesaugustobriga.juntaextremadura.net/memoria/Dichos.htm
- Spanish Proverbs – Wikiquote - http://en.wikiquote.org/wiki/Spanish_proverbs
- Lista de paremias en español (Centro Virtual Cervantes) - http://cvc.cervantes.es/lengua/refranero/listado.aspx

Conclusions

Most of the data sources (NLP resources and images) are generally freely available for academic, research and non-commercial projects, but require additional licensing for commercial exploitation. As long as FIRST is a research project, fair use of such resources can be claimed and issues are not expected to arise.

For the commercial product we are developing we identified the licensing of each NLP resource used for testing end development and checked which allows us to use them. Some of the resources can be used freely for commercial purposes, others only after certain conditions/requirements are met. We are not allowed to use the rest of the resources commercially.

OpenBook will rely on external image libraries. At the moment, images returned from Google, Bing and Wikipedia search engines are inserted in the text to explain the meaning of a word.

Both search engines (Google and Bing) place limitations on a number of queries that can be done in free mode (for Google the limit is 50 queries/day, for Bing there is a monthly limit). To overcome this issue twenty accounts were created for testing purposes (10 for each site). The created clients were implemented for the services that allow image retrieval for almost every concept in English, Spanish and Bulgarian.

The purpose of the created Google and Bing accounts is to enable programmers to access Google and Bing data. In fact it is the only way in which the availability of images for almost every term can be guaranteed. It is free to create accounts and they can be integrated in the actual web service.

It is envisaged that OpenBook users register for Google and Bing free accounts (for the final version of the tool demo movie about the registration process can be recorded). The registration process for one user takes approximately 20 minutes. The accounts keys should then be stored in the user profiles and passed to the Web Service in the appropriate way.

Since there are no critical issues for the FIRST project, we foresee the following challenges for the commercial realisation of OpenBook.

Problems for the commercial product:

1. Some of the NLP resources can be used only for research purposes (the TreeTagger, Babelnet, Bulgarian National Corpus, Bulgarian Treebank word frequency list, etc). Their commercial use is not allowed. We should consider alternative solutions for the usage of such resources or contact the copyright holders, where available, for license agreements.
2. The problem with image URLs is complicated by the fact that images can be located on servers all over the world, and thus be subject to different local copyright legislation. Such images cannot be made publicly available on one server. For example, Google and Microsoft image search engines will not be appropriate because they show images from various locations.
3. Internet image protection might be a serious problem. That's why some of the image providers who aim to keep their rights only provide pictures with watermarks. Such pictures can't be used without a license.
4. Some of the pictures displayed by the image retrieval service may be inappropriate for people with ASD, especially for children.

Solutions:

A universal solution for the image retrieval service is to use link to an image from Google/Bing instead of the image itself. If the user is not happy with any of the images suggested by the system, he can add his own resource. The system will prompt for a URL to the image resource and will automatically insert the image into the text, also notifying OpenBook Integration Engine that an image resource has been used for the selected word. Users will not be allowed to upload their images to OpenBook, because that might cause copyright issues and put a heavy strain on the hardware requirements of the servers. Therefore, users will be required to upload their pictures on an image hosting service, and provide a link.

Other solutions include:

1. To drop the image and dictionary libraries and have the potential customers integrate their own resources into them.

2. To approach a big image provider and see whether they would be willing to provide the images for our tool (for free or for costs). Companies like Getty Images would probably be appropriate because they either hold the copyright or they have an agreement with the copyright holders in order to distribute images.
3. Identify sites that do not have copyright restrictions and use our Google and Bing clients only with these sites.
4. To create a local database with images that have no copyright restrictions. That would take a serious amount of work.

At this stage, all the resources necessary for the commercial product Open Book are still not certainly defined. Detailed research for the rest of the required licenses will be carried out next year.

The copyright holders, where available, will be contacted for license agreements. We are prepared to analyse the outcome of the negotiations with copyright holders, and we foresee the following options:

1. Purchase the required license.
2. Payment of a fee based on the commercial turnover of the project.
3. Excluding resources from the project or providing only a limited free resource. B2B customers will be allowed to provide their own resources to improve the level of the text simplification service.

Picking a solution would become possible when we have achieved tangible results with OpenBook and we have the feedback from actual users. Only then it would be possible to estimate the commercial potential of OpenBook.

Appendix

Licenses of NLP Resources

Licensing Scheme	Software	Research	Commer-cial
Creative Commons Attribution 3.0 Unported License	The Paraphase Database, MultiWordNet, Wiktionary, WordNet Domains.	✓	✓
Creative Commons Attribution-NonCommercial-Share Alike 3.0 License	Babelnet, Bulgarian National Corpus (BulNC), Wikipedia Images	✓	X
MS-NC-NoReD (META-SHARE Non Commercial No Redistribution) License Agreement	Bulgarian Treebank word frequency list	✓	X
GNU General Public License	Freeling 3.0, Weka, BalkaNet	✓	✓
GNU Lesser General Public License	Gate 7.0, JWPL	✓	✓
Common Public License	MALLET	✓	✓
BSD 3-Clause License	JWNL 1.4.1	✓	✓
Apache License, Version 2.0	Apache Lucene TM	✓	✓

GNU Free Documentation License	Wiktionary	✓	✓
Software licensed under the Linguistic Data Consortium License	Coreference Resolution in Multiple Languages, English and Spanish Word Sense Disambiguation, Multilingual Word sense disambiguation task	✓	X
Software licensed by Universitaet Stuttgart	TreeTagger	✓	X
SIL Open Font License 1.1	Font Awesome	✓	✓
ImageNet Licensing Agreement	ImageNet	✓	X
Oxford Text Archive License	Kucera and Francis frequency list	✓	X