Semantic Annotation and Classification in Practice: Use Cases of a Telecommunication Operator

Oscar Rodrìguez Rocha[1], Iacopo Vagliano[1], Cristhian Figueroa[1], Federico Cairo[2], Giuseppe Futia[2], Carlo Licciardi[3], Marco Marengo[3] and Federico Morando[2]

Politecnico di Torino, corso Duca degli Abruzzi 24, 10129. Turin, Italy.
{oscar.rodriguezrocha,iacopo.vagliano,cristhian.figueroa}@polito.it
Nexa Center for Internet & Society, via Pier Carlo Boggio, 65, 10138, Turin, Italy.
{federico.cairo,giuseppe.futia,federico.morando}@polito.it
Telecom Italia, via G. Reiss Romoli, 274, 10148. Turin, Italy.
{carlo.licciardi,marco.marengo}@telecomitalia.it

Abstract. The evolution of the traditional Web into a Semantic Web and the continuous increase in the amount of data published as Linked Data open up new opportunities for annotation and categorization systems to reuse these data as semantic knowledge bases. Accordingly, Linked Data has been used by information extraction systems to exploit the semantic knowledge bases, which can be interconnected and structured in order to increase the precision and recall of annotation and categorization mechanisms. This paper describes TellMeFirst a software for the classification and enrichment of textual documents written in English and Italian. Although nowadays there are various works presenting solutions for text annotation and classification, this work is focused on describing and studying the use case of a Telecommunications Operator that has adopted TellMeFirst in order to generate value-added to two services available to its users: FriendTV and SOCIETY.

Keywords: Information Extraction, Linked Data, Semantic Annotation, Semantic Classification

1 Introduction

In 2011, the Nexa Center for Internet & Society of the Department of Computer Science and Automation of the Politecnico di Torino (Italy) started a project to create a software tool for the automatic classification and enrichment of documents named TellMeFirst. It is available under the GNU AGPLv3 license at GitHub (see http://github.com/TellMeFirst).

This article presents an overview of TellMeFirst making emphasis on its main features such as the semantic annotation and classification. Although some technical details have been included to give more information to those readers interested in learning about the inner workings of the system, the main goal of this article is to show how this software has been used on two practical industrial cases by Telecom Italia, one of the major Telecommunications Operators in Italy.

Such operator, looking for ways to add value to its services, has decided to introduce functionalities provided by TellMeFirst. The first one is Friend TV, a social television service of that allows users to share television experiences with other viewers on the social media through tablets and smartphones. The second one is SOCIETY, a platform that allows end-users to share notes and comments with other users in a social community while reading an ebook.

In conclusion, this article seeks to give the reader a concrete example how research and innovation may provide advantages at business level, by being applied in real commercial services.

2 Background

The goal of the Semantic Web is to describe the meaning of the information published on the Web to allow retrieval based on an accurate understanding of its semantics. The Semantic Web adds structure to the

resources accessible online in ways that are not only usable by humans, but also by software agents that can rapidly process them [1].

Linked Data (LD) refers to a way for publishing and interlinking structured data on the Web (the Web of Data). LD is part of the design of the Semantic Web and represents the foundation brick needed to build it. Although the concept of LD was already present in the first theory of the Semantic Web [3], it came into vogue later in computer science. However, in the recent years due to the growing number of datasets based on LD, it has been possible to exploit their implicit knowledge through text classification and annotation processes to build semantic applications.

Text classification is the assignment of a text into one or more pre-existing classes (also known as features). This process determines the class membership of a text document given a set of distinct classes with a profile and a number of features [14]. The criterion for the selection of relevant features for classification is essential and is determined a priori by the classifier (human or software). The semantic classification takes place when the elements of interest in the classification refer to the meaning of the document.

Text annotation refers to the common practice of adding information to the text itself through underlining, notes, comments, tags or links. The annotation of text can be also semantic when the text of a document is added with information about its meaning or the meaning of individual elements that compose it [13]. This is done primarily using links that connect a word, an expression or a phrase to an information resource on the Web or to an unambiguous entity present in a knowledge base [8].

3 TellMeFirst

The TellMeFirst project was started in October 2011, thanks to the funding of the *Working Capital - National Innovation Award*. It was mainly developed within the Nexa Center for Internet & Society of the Department of Computer Science and Automation of the Politecnico di Torino (Italy).

TellMeFirst is a software tool for the automatic classification and enrichment of documents that uses DBpedia (the Linked Data version of Wikipedia) as the reference knowledge base for content extraction and disambiguation, in a similar way as other software tools of the same kind (such as DBpedia Spotlight and Apache Stanbol). DBpedia has been chosen for the semantic classification because the Wikipedia corpus is a perfect training set for every categorization approach based on Machine Learning (an approach which allow to learn from data [9]), and for the semantic annotation because of its direct connection with the vast multilingual pre-annotated corpus of Wikipedia [10].

As noted, TellMeFirst takes advantage of the relationship between Wikipedia and its semantic counterpart DBpedia to perform its semantic annotation and classification processes in a quick and efficient manner. While this feature distinguishes it from other similar tools, also makes it dependent on these datasets. Given the open nature of the Web of Data, which is not limited to a single dataset, it is important to consider a future evolution towards compatibility with multiple datasets.

3.1 Semantic Annotation

The semantic annotation process of TellMeFirst consists in associating semantic information to the words contained in a text, i.e. identifying which meaning of a word is used in a sentence.

This problem is well known as word-sense disambiguation (WSD). To address this problem, TellMeFirst provides a *disambiguator* that implements three sub-components: a Knowledge-based Disambiguator, a Corpus-based Disambiguator and a First Sense Heuristic Disambiguator. The latter comes into operation when the former does not have supplied a result accompanied by a sufficient level of confidence. It is a mechanism that exploits the *coeficient of prominence* of Wikipedia resources, or the number of times in which they are mentioned in Wikipedia through a *wikilink*, to decide on the most common meaning of an ambiguous term. When a term is not disambiguated by the Knowledge-based Disambiguator and from Corpus-based Disambiguator with a certain degree of confidence, then the First Sense Heuristic Disambiguator assigns the most common meaning. The

heuristic approach has been shown to be often only a few percentage points below the WSD systems with higher performance [11].

Table 1 summarizes the test results of the TellMeFirst annotators, carried out on a corpus of 10 excerpts from a newspaper. The last column shows the possible usage scenarios of the different disambiguators.

Table 1. Tests on the TellMeFirst's Disambiguator

Disambiguator	Average time per word[s]	Average precision	Average recall	Canonical use case
Corpus-based	0,04	0,85	0,21	Online annotation of news portals or blog
Knowledge-based	0,07	0,99	0,05	Automatic classification of documents based on DB-pedia
First sense heuristic	0,04	0,78	0,24	Online annotation of news portals or blog in a more generic boundary, where the most common Wikipedia meaning is the most likely
Default	0,10	0,96	0,08	Offline annotation, automatic classification, text enhancement

3.2 Semantic Classification

TellMeFirst implements a memory-based learning approach to semantic classification, a subcategory of lazy learning [6]. A distinctive feature of this approach, also known as instance-based learning, is that the system does not create an abstract model of the classification categories (profiles) before the process of text categorization. Instead, it assigns the target document to a class on the basis of a local comparison between the pre-classified documents and the target [4,5].

The classifier must hold in memory all instances of the training set and calculate, during the classification stage, the distance vector between the training documents and the not classified ones. This approach belongs to the family of lazy learning, which refers to the classification phase (consultation time) and the calculation of the similarity with the training set. Another kind of approach is the eager learning, which anticipate this operation to the learning phase (training time) where the specific profiles of categories are created and the function to perform the classification is defined [12].

The semantic classification processes of TellMeFirst is performed by the *k-Nearest Neighbor* algorithm (kNN). This algorithm is a type of memory-based approach which chooses the category or categories, to which the target document belongs, based on the k most similar documents to a target in a space vector [12].

The training set consists of all the paragraphs in which there is a link within Wikipedia (wikilink). These paragraphs are stored in an Apache Lucene index as *Field CONTEXT* of documents that represent DBpedia resources. In this index, each DBpedia resource (and subsequently each Wikipedia page) corresponds to a Lucene Document, and for each Document there are as many Field CONTEXT as paragraphs in which the resource appears as wikilink.

In the classification (following a lazy approach) the target document is transformed into a Boolean Lucene query over the Field CONTEXT of the index, to discover the conceptual similarity with the contexts of Wikipedia entries. To calculate the similarity, the Lucene's Default Similarity is used, which combines the Boolean model with the Vector Space Model (VSM).

The results approved by the Boolean search on the index, are then sorted according to the VSM. Lucene takes care of the *stemming*, lemmatization and the filtering (through specific stop words either for Italian or English) of the features of both the training documents and the target document transformed into a query. The query and the training documents become both feature vectors (depending on the model of the *bag of words*), where the weight of each feature is calculated according to the TF-IDF (Term Frequency Inverse Document Frequency) algorithm. The query returns a list of documents(DBpedia URIs) ordered according to a *similarity score* which is based on the *cosine-similarity*. Cosine-similarity is a well-known metric that has proven to be robust for scoring the similarity between two textual strings, and is frequently used in complex queries [2]. Once the ordered list of results is obtained, the *RCut* method is applied for thresholding [15,7], keeping only the first 7 results according to their *rank* and discarding the others.

The technique used by TellMeFirst for the classification is based on the VSM for the representation of both the training documents and the target document (it may be considered of the spatial type). The similarity between two documents can be viewed geometrically as the distance between the two vectors that represent the documents in a n-dimensional vector space, where n is the number of features of the entire corpus of training. The vector space model is also the basis of the Lucene libraries. Lucene is optimized to quickly perform the calculation of the distance between the documents according to the TF-IDF algorithm: given a query that represents the features of the target document, it returns a list of similar documents indexed, even when the index is populated by millions of documents. The score obtained with Lucene represents the inverse of the distance between two documents: the higher the score, the closer are the documents in the vector space.

To show the results, TellMeFirst provides a visualizer that contains a window with 7 frames of diverse size (according to the first 7 results previously ranked). Each frame indicates an argument extracted from the text and its size represents the relevance of the text. Figure 1 shows an example of the results displayed in the 7 frames of the TellMeFirst visualizer (other examples can be run on the available demo at: `http://tellmefirst.polito.it`).

The next section describes the two services, of the Telecommunications Operator, enhanced by TellMeFirst.

Fig. 1. Example of the results displayed by the TellMeFirst visualizer

4 Society

SOCIETY is a platform of Telecom Italia that allows end-users to share notes and comments, while reading an ebook, with other users in a social community. In this service TellMeFirst makes it possible to analyze these contents to extract semantic concepts and hence to enable readers to deepen the information contained.

SOCIETY (available as mobile applications at `https://play.google.com/store/apps/details?id=it.telecomitalia.society` for android and `https://itunes.apple.com/it/app/society-school-2.0/id785451519?mt=8` for iOS) is composed by a community of readers, able to share comments over a paragraph or even contributing to improve the ebook by sending correction reports to the authors. Groups of readers can be built based on social network relationships. The comments over a reading text can be shared as notes into the social network, thus propagating to other

users based on a configuration. Each note can be shared by the users through a specific interface to the most used social network such us Facebook or Twitter. Through this interface, friends or followers can see what other users did, read their notes and add comments or retweet the note giving more results to this piece of information. A specific interface from the social network to the system node platform could be provided in order to extract and enrich notes with information provided by users on the social network platforms (Figure 2).

One of the most important features is integration with TellMeFirst in order to semantically annotate the user-generated notes. When creating a new note or a comment on SOCIETY, TellMeFirst analyzes them in order to recognize each relevant entity such as places, names, concepts and links them with concepts or resources present in the Web of Data. The main results could be returned and showed by the application to the user interface giving to the user the possibility to save it as a note that add extra information to the book.

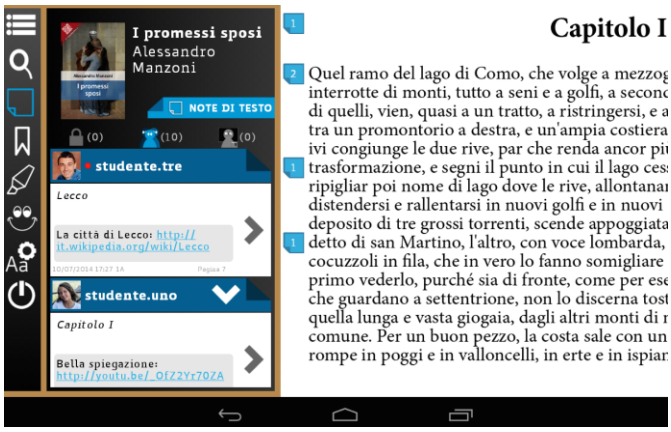

Fig. 2. The graphic interface of the Society application for Android devices

The SOCIETY application can be used on-the-go, as it is aware of the user's context (detecting an entity as a place). In this case, the semantic source can also provide localization information and be used as extra fields for searching more contents such as multimedia user generated contents that match the same localization.

SOCIETY also provides the traditional search in order to add extra information using the World Wide Web as a common source for multimedia and extra information, such as image, video, audio and text information related to words or sentences written in the book. At the same time, the note platform can provide its information to other applications in order to show note in their target interface based on localization information. Moreover, the same application provides some accessibility functions, which extend the book-reading experiences also to the people with limited capacities such as blindness, by reading the text through a text to speech (TTS) engine and low vision, by adjusting the size of the fonts.

This application fits perfectly into education initiatives aiming to schools digitalization, allowing the interactive education processes to avoid the hard-printed books and making possible the exchange of instant messages (alternative to chat) between teacher and students.

At the moment, this service is being exploited by Telecom Italia mainly as a social effect initiative aimed to support persons with limited capacity and improve the efficiency of the education process in schools. However, an economic income is also possible through eBooks distribution supporting this social comments exchange features.

5 FriendTV

FriendTV is a social television service of Telecom Italia that allows users to share television experiences with other viewers on the social media through tablets and smartphones. FriendTV presents a list of TV programs that may be of interest to an end-user. This service uses a semantic annotator and classifier provided by

TellMeFirst in order to extract and associate the concepts (based on their semantic meaning) contained in the description of each program with related existing web resources. Hence, the user can easily browse for additional information about these related concepts.

This service can be seen as a television guide integrated with Twitter and Facebook, where users can discuss about the most followed television programs on these social media and also to receive suggestions about them. In FriendTV a user can obtain information about the scheduled television programs, communicate to other users what he is watching and set notification about television programs he is interested in, in order to be aware when they are broadcasted. He can also rate television programs enabling the system to provide better recommendations. In addition, the service allows broadcasters and media agency to release questionnaire, compute statistics based on social media and insert banners for giving information about programs or advertising.

TellMeFirst is integrated in this service in order to provide to the users the content related to a program. In fact, it is possible to open a detailed view on a specific program and receive related content, e.g. related videos. Thus, by starting from the TV program description, TellMeFirst is able to annotate the text, classify it and exploit the links with other resources generated by the annotation in order to retrieve semantically related content. The workflow of the FriendTV service is depicted in figure 3: (1)the user select a program. (2) the raw text composing the description of the selected program is given as input to TellMeFirst. (3) TellMeFirst generates the annotated version of same text, in which a number of entities are linked to existing and related web resources. (4) These resources are provided to the user as related content.

More information about FriendTV can be found at `http://www.stv.telecomitalia.it/`. Additionally, it is available as mobile application for Android (`https://play.google.com/store/apps/details?id=it.telecomitalia.friendtv`) and iOS (`https://itunes.apple.com/it/app/friendtv/id784514746`) devices. At the moment it counts with thousands of downloads.

6 Conclusions

The Web is facing a crucial challenge to promote the construction of a new knowledge infrastructure. This is one of the fundamental tasks of the Linked Data in order to achieve the vision of a Semantic Web. The presented software platform (a research and development project) takes advantage of the information present in the Web of Data to generate semantic annotation and classification of concepts. It is noteworthy that in order to give the maximum benefit from this platform to end users, it was necessary to establish a joint collaboration with the main Telecommunications Operator in Italy (which in our case, provides mobile services), and in this way to put into direct contact common users of mobile services with the Web of Data.

When users use these services in their real life, both the creators of TellMeFirst and the Telecommunications Operator are able to obtain benefits. In fact, through the increased use of the functionalities of semantic annotation and classification, it has been possible to detect improvement areas. In addition, with these new features, the operator can think of new ways to monetize these services and make them increasingly innovative.

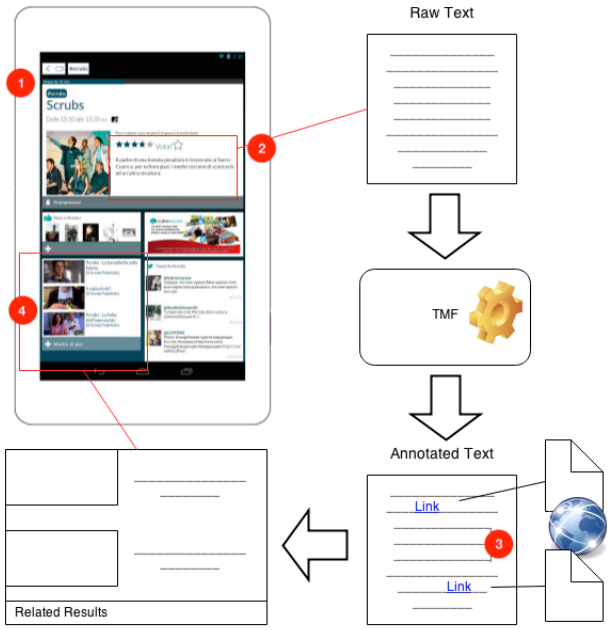

Fig. 3. Use of TellMeFirst in FriendTV

As future work, it is planned to improve the functionalities of TellMeFirst by introducing a Linked Data based Concept Recommender able to suggest similar concepts related to those originally extracted in the initial semantic annotation process. This improvement will enable a whole new scenario of multi-domain recommendations. Additionally, as highlighted in section 3, many efforts are being made to adapt the operation of TellMeFirst with multiple knowledge bases (and not just DBpedia as currently).

7 Acknowledgements

This paper and the services described are the result of a joint cooperation between the Software Engineering Research Group and the Nexa Center for Internet & Society at Politecnico di Torino and Telecom Italia. We thank everyone who contributed to the design and creation of these services.

References

1. H.P. Alesso and C.F. Smith. Thinking on the Web: Berners-Lee, Gödel, and Turing.2009.
2. D.C. Anastasiu and G. Karypis. L2ap: Fast cosine similarity search with prefix l-2 norm bounds. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 784-795, March 2014.
3. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28-37, 2001.
4. Weiwei Cheng and Eyke Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3):211-225, 2009.
5. Bin Fu, Zhihai Wang, Guandong Xu, and Longbing Cao. Multi-label learning based on iterative label propagation over graph. *Pattern Recognition Letters*, 42(0):85 -90, 2014.
6. Aastha Gupta, Rachna Rajput, Richa Gupta, and Monika Arora. Hybrid model to improve time complexity of words search in pos tagging, Sept 2014.
7. Xiaofeng He, Rong Zhang, and Aoying Zhou. Threshold selection for classification with skewed class distribution. In Yunjun Gao, Kyuseok Shim, Zhiming Ding, Peiquan Jin, Zujie Ren, Yingyuan Xiao, An Liu, and Shaojie Qiao, editors, *Web-Age Information Management*
, volume 7901 of Lecture Notes in Computer Science, pages 383-393. Springer Berlin Heidelberg, 2013.

8. Alexander Hogenboom, Frederik Hogenboom, Flavius Frasincar, Kim Schouten, and Otto van der Meer. Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications*, 64(1):27-52, 2013.
9. Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30:271-274, 1998.
10. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 2014.
11. Diana McCarthy. Word sense disambiguation: An overview. *Language and Linguistics Compass*, 3(2):537-558, 2009.
12. Claude Sammut and Geoffrey I. Webb, editors. *Encyclopedia of Machine Learning*. Springer, 2010.
13. David Sánchez, David Isern, and Miquel Millan. Content annotation for the semantic web: an automatic web-based approach. *Knowledge and Information Systems*, 27(3):393-418, 2011.
14. V.K. Singh, R. Piryani, A. Uddin, P. Waila, and Marisha. Sentiment analysis of textual reviews; evaluating machine learning, unsupervised and sentiwordnet approaches. In *Knowledge and Smart Technology (KST), 2013 5th International Conference on*, pages 122-127, Jan 2013.
15. Yiming Yang. A study of thresholding strategies for text categorization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 137-145, New York, NY, USA, 2001. ACM.

Oscar Rodrîguez Rocha holds a PhD in Information and Systems Engineering from the Politecnico di Torino. He's mainly interested in studying how Semantic Web (Linked Data), Context Awareness and User-Generated Content can improve Mobile Services. He has contributed to European research projects and collaborated with the Telecom Italias Context Awareness research group to design and prototype Mobile Services. He holds a second level Master degree in Wireless Systems and Related Technologies from the Politecnico di Torino and graduated from Universidad Panamericana in Mexico City as Computing Systems Engineer with an specialization in Computing Networks and Distributed Systems.

Iacopo Vagliano is a Ph.D. student at the Politecnico di Torino in Computer and System Engineering. He graduated in same university in Computer Engineering. He is currently working with Linked Data and other Semantic Web technologes, addressing recommender systems and data integration. He is collaborating with Telecom Italia in the JOL MobiLAB.

Cristhian Figueroa is Ph.D student at Politecnico di Torino in Computer and Systems Engineering. He is a master in Telematics at University of Cauca (Colombia) and Electronic and Telecommunications Engineering at Univeristy of Cauca. His research topics are mainly Semantic Web, Linked Data, Mobile Information Systems, Web Services Retrieval, and Business Process retrieval. He is currently working with Recommender Systems based on Linked Data.

Federico Cairo, Ph.D., is a project manager at Expert System SpA and a fellow at the Nexa Center for Internet & Society at Polytechnic University of Turin. His main research interests are Linked Data technologies and natural language processing. He is the technical lead of TellMeFirst, an open-source software for classifying and enriching textual documents via Linked Open Data.

Giuseppe Futia joined the Nexa Center for Internet & Society in February 2011. He holds a Master Degree in Cinema and Media Engineering in 2011 and from 2008 he collaborates regularly with the Italian newspaper "La Stampa". At the Nexa Center, he is in charge of communication and media management. Giuseppe also takes part in several ICT projects, dealing with different technical tasks, including support for management of cloud computing platforms (Open-DAI) and development of Natural Language Processing and Linked Data

applications (TellMeFirst) within Internet Science studies (EINS). Giuseppe holds data visualization skills, useful to both sustain the outreach of some of the Nexa projects, and to support research in the field of Open Data.

Carlo Alberto Licciardi is Technical Manager at Telecom Italia , Strategy and Innovation department. He joined Telecom Italia Lab (formerly CSELT -Torino, Italy, leader company in research and development for Telecommunication) in 1992 where he has worked in long term aspects of Intelligent Network and in the design of software architecture for the provisioning of Advanced Telecommunication Services. He has contributed to standardization activities (ITU-T, 3GPP, JAINSLEE and OMA) and to worldwide research projects (IST, TINA-C and EURESCOM). He has been project leader of several European projects on evolution of TELCO Service

Layer. He has been steering board member in EU projects dealing with Context Awareness, Semantic web, Cloud based software architecture and Social Networking: CCAST (Context aware content casting), MUSIC (Middleware for context aware adaptation), BUTLER, FI-ware. He is currently leading internal research project in Telecom Italia in the area of Mobile social communities enablers and Context Awareness, dealing with the delivery of seamless Mobile communities context aware services which support advance content and communication services for mobile and fixed Telecom Italia customer. His current research interest is in definition of context representation languages, Social Big data platform and semantic analysis of social stream, user profiling and advanced service adaptation for mobile and desktop customers. He is author of several scientific papers in the area of service creation, application server for next generation networks and context awareness.

Marco Marengo is a Computer Engineer, graduated at Politecnico di Torino in 2004. Since then, he has been a researcher in Telecom Italia. At the beginning he worked on a VAS platform running on StarSIP, then he devoted himself to designing and creating mobile applications. He has worked with the main mobile platforms: Windows Mobile, Windows Phone, Android e iOS, and last but not least HTML5. He love technology, fantasy books, sports and photography. He published three photographic ebooks related to volleyball. Since 2013 he is the Director of JOL MobiLAB, and he is leading a young and highly motivated team.

Federico Morando is an economist, with interdisciplinary research interests focused on the intersection between law, economics and technology. His research activity at the NexaCenter mainly concerns new models of production and sharing of digital contents. He also taught intellectual property and competition law at Bocconi University in Milan and he is an associate editor of the IJCLP. He has an undergraduate degree in Economics from Bocconi Univ. and a masters degree in Economic theory and econometrics from the Univ. of Toulouse. He holds a Ph.D. in Institutions, Economics and Law from the Univ. of Turin and Ghent with a dissertation about software interoperability. He joined the working group of the Nexa Center in 2006 at the beginning of its first year of formal activity. From Dec. 2012, he leads the Creative Commons Italy project and he is a member of the Open Team of Regione Piemonte that launched and steers the development of the first Italian open government data portal. From 2008 to 2012, in his position as the first Managing Director of the Center, he worked closely with the Directors to define staff and project goals and to coordinate the Centers fellows. From 2013 he is Director of Research and Policy, focusing on the coordination of the cross-disciplinary research activities of the Center and on the related policy support actions.