The KNOWLEDGESTORE: an Entity-Based Storage System

R. Cattoni, F. Corcoglioniti, C. Girardi, B. Magnini, L. Serafini, R. Zanoli

Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
{cattoni,corcoglio,cgirardi,magnini,serafini,zanoli}@fbk.eu

Abstract

This paper describes the KNOWLEDGESTORE, a large-scale infrastructure for the combined storage and interlinking of multimedia resources and ontological knowledge. Information in the KNOWLEDGESTORE is organized around *entities*, such as persons, organizations and locations. The system allows (i) to import *background knowledge* about entities, in form of annotated RDF triples; (ii) to associate resources to entities by automatically recognizing, coreferring and linking *mentions* of named entities; and (iii) to derive new entities based on knowledge extracted from mentions. The KNOWLEDGESTORE builds on state of art technologies for language processing, including document tagging, named entity extraction and cross-document coreference. Its design provides for a tight integration of linguistic and semantic features, and eases the further processing of information by explicitly representing the *contexts* where knowledge and mentions are valid or relevant. We describe the system and report about the creation of a large-scale KNOWLEDGESTORE instance for storing and integrating multimedia contents and background knowledge relevant to the Italian Trentino region.

Keywords: LR Infrastructures and Architectures, Knowledge Discovery/Representation, Metadata

1. Introduction

In the digital World, relevant knowledge about entities of interest – such as persons, organizations and places – may be contained both in structured and unstructured resources. Examples of structured resources are databases and ontologies; unstructured resources are textual documents, images, videos and multimedia documents in general.

Decades of research and technological advancements have led to the development of models, systems and best practices to manage structured and unstructured information resources, but this has mainly happened by considering each kind of resource in isolation. Given the exponential increase of multimedia digital documents and the rise of the so-called *Web of Data* – mainly consisting of interlinked RDF (Carroll and Klyne, 2004) datasets published as part of the Linked Data initiative (Heath and Bizer, 2011) – the need arises for an advanced content management system supporting the joint management of structured and unstructured information resources. Above all, such a system should support the interlinking of the two kinds of resource, which can be obtained by (i) identifying mentions of named entities in multimedia resources, (ii) establishing coreference among such mentions (i.e. cross-document coreference) and (iii) linking entity mentions to corresponding structured entity descriptions.

In our vision, this advanced content management system should be inspired by the following principles:

- *Scalability*. As large multimedia collections and structured datasets are becoming widespread and publicly accessible on the Web, scalability with respect to the size of managed data is a crucial matter.
- *Traceability*. Through the use of rich metadata, stored information should be traced back to its location in the original information sources, so to guarantee the proper use and exploitation of information contents.
- *Reproducibility*. The processes managing the interlinking of information resources should be automatic and reproducible, given the initial raw resources.
- *Incrementality*. At any moment, it should be possible to add new information sources (or remove existing ones) and rely to the system for the proper merging and interlinking of new contents with existing ones, without the need to re-process all stored information.
- *Contextualization*. As stored information is generally valid or relevant only in certain contexts and not universally, these contexts should be explicitly represented and associated to information elements.

Based on that vision, in this paper we present the current results of our ongoing work in developing the KNOWLEDGESTORE, a system intended to fill the current gap in content management systems. The KNOWLEDGESTORE combines a scalable storage infrastructure, able to store both multimedia resources and structured knowledge in form of annotated RDF entity descriptions, with state-of-the-art systems for document tagging, cross-document coreference and mention-entity linking. We describe the KNOWLEDGESTORE as a general-purpose system and then we focus on reporting our experience in creating a specific large-scale KNOWLEDGESTORE instance in the scope of the *LiveMemories* project[1]. In this context, we exploited the KNOWLEDGESTORE to store and interlink large amounts of multimedia documents and ontological knowledge about the Italian Trentino region, with the aim of realizing a comprehensive repository of the knowledge and the digital memories of this area.

The remainder of the paper is structured as follows. Section 2 presents the KNOWLEDGESTORE system, with its data model (section 2.1) and the processing pipeline used to elaborate and interlink information contents (section 2.2). Section 3 presents the Trentino KNOWLEDGESTORE instance, with subsections describing step by step how we populated the system and processed the data. Section 4 reports on related works and section 5 concludes.

[1] www.livememories.org

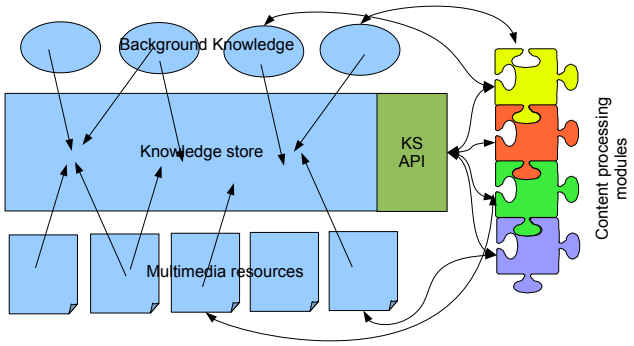

Figure 1: The architecture of the KNOWLEDGESTORE system: the KNOWLEDGESTORE *core* contains both Background Knowledge and information extracted from multimedia data by the Content Processing modules. Such modules interact with the KNOWLEDGESTORE *core* by means of web services provided by the KNOWLEDGESTORE API.

2. The KNOWLEDGESTORE System

The architecture of the KNOWLEDGESTORE system is shown in Figure 1: it is composed by the storing infrastructure – the KNOWLEDGESTORE *core* – and a *pipeline* composed of *content processing modules* interacting with the *core* through a specific API. The KNOWLEDGESTORE *core* is populated with several types of information: the original multimedia resources, the pre-existing background knowledge, and the data extracted from the multimedia resources by the content processing modules. It is worth noticing here that content processing modules may utilize the background knowledge to perform their tasks.

Concerning the implementation, the KNOWLEDGESTORE *core* is realized on top of the Hadoop[2] and Hbase[3] frameworks. The API interacting with the content processing modules consists of a set of Web services realized with Tomcat servlets.

The remaining of the section first describes the data model adopted by the KNOWLEDGESTORE to represent and store the data (section 2.1). Then, the processing pipeline that populates the KNOWLEDGESTORE is presented with some details (section 2.2).

2.1. Data Model

The data model of the KNOWLEDGESTORE is shown in the UML class diagram of figure 2. Information in the KNOWLEDGESTORE is represented at four interconnected layers: resources, mentions, entities and contexts.

Resources. This is the layer where physical files and their metadata are stored. Examples are multimedia data such as texts, images, audios and videos (or portions of them). Also the files derived from processing of original data such as Automatic Speech Recognition transcriptions and linguistic annotations are resources. Metadata of a resource are represented with an object named *Media*. The reference for metadata description is the *Dublin Core Metadata Standard*[4]. As figure 2 highlights, the Media class includes additional attributes introduced to represent relations among resources: `captionOf`, `from`, `partOf` and `relatedTo`. While the first three attributes encode specific and self-explaining relations, `relatedTo` is utilized for a generic relation between resources.

Mentions. This layer stores portions of a resource referring (i.e. mentioning) to a relevant content. Mentions may refer to persons (PER), organizations (ORG) and geo-political and location entities (GPE/LOC). Mentions can appear in medias of different types. For instance, a textual PER mention is a fragment of text (e.g. "President Mario Monti"), while a mention of the same person in a picture is represented as the area of the picture where Mario Monti is depicted. We have taken advantage of the definitions provided in the context of the ACE (Automatic Content Extraction) program[5]. Mentions are a kind of hybrid object as on the one side they are portions of the original resource, while, on the other side, they are enriched with semantic information that we extract and store in the KNOWLEDGESTORE. As an example, for the mention "President Mario Monti" we may store the fact that this mention represents a person, whose `firstName` is "Mario", whose `lastName` is "Monti" and whose `activity` is "President". Semantic information for PER mentions is mainly extracted from certain kinds of words – called `triggers` – that appear immediately before or after the mention (such as "President" in previous example). Mentions are defined in the context of the document they occur in, and, as a consequence, their semantic annotations hold only in that precise context.

Entities. This is the layer storing unique individuals that are referred through mentions occurring in resources. Entities are described through ⟨`subject`, `predicate`, `object`⟩ triples (the subject being always the entity), which may derive both from automatically recognized mentions and from pre-existing *background knowledge* loaded into the KNOWLEDGESTORE. Under this view, entities are supposed to store information about individuals, such that multiple occurrences of the same information are represented in a single triple, no matter how many variants are used to express it in documents. Typically, there is a many-to-one relation between mentions and entities, as the same entity can have several mentions, in different documents. The two mechanisms that allow to move from mentions to entities are *coreference* and *linking*: as an example, the two mentions "President Mario Monti" and "President Monti", although different, are clustered together and linked to the same entity MARIO MONTI. As figure 2 points out, the Triple class includes additional attributes encoding triple metadata. In particular, attribute `source` stores the provenance of a triple, while attributes from `crystallized` to `compatibility` have been added in the perspective of extending the KNOWLEDGESTORE with a process for consolidating knowledge extracted from mentions into triples.

Contexts. A piece of knowledge always holds in a certain context. For instance, the fact that Mario Monti is President only holds in the context of a specific time period, and it is not true in other circumstances. Within the

[2]http://hadoop.apache.org
[3]http://hbase.apache.org
[4]http://dublincore.org/documents/dces/
[5]http://www.itl.nist.gov/iad/mig/tests/ace/

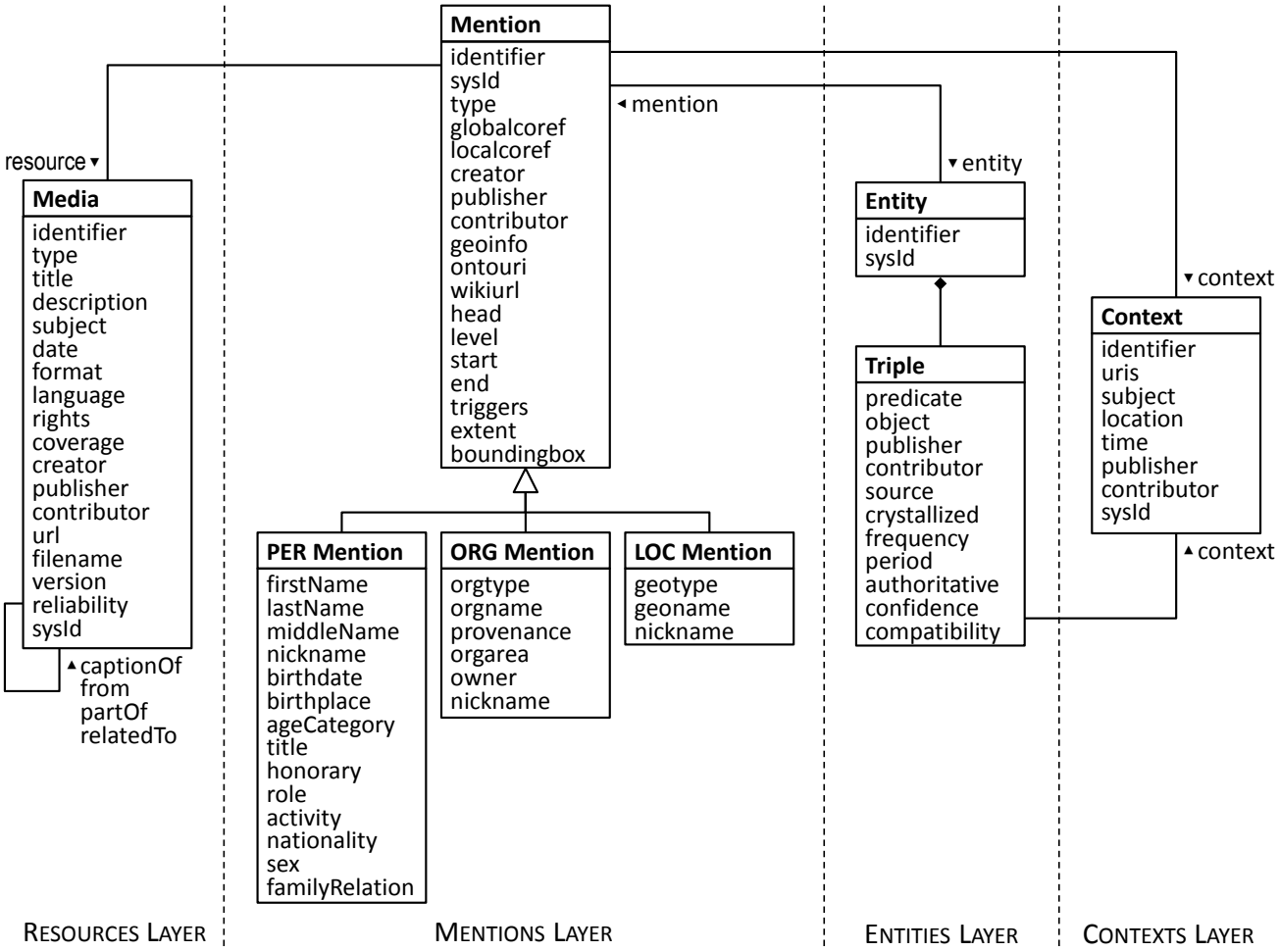

Figure 2: The KNOWLEDGESTORE data model.

KNOWLEDGESTORE we have formally defined a hierarchy of contexts (from very general to very specific ones), where knowledge about entities is supposed to be true. Following the CKR approach to model contextualized knowledge (Serafini and Homola, 2011), a contexts is identified with a ⟨`subject`, `location`, `time`⟩ tuple. The values of the subject, location and time dimensions are organized by means of *cover* relations that, jointly, permit to classify contexts in a broader-narrower hierarchy, e.g. to classify context ⟨`Sport`, `Italy`, `2008-2012`⟩ as being broader in scope than context ⟨`Football`, `SerieA`, `2010`⟩. Contexts may be induced either from the textual contexts of mentions or from pre-existing background knowledge.

2.2. Processing Pipeline

A KNOWLEDGESTORE instance is initially fed with raw multimedia resources and background knowledge, which provide the initial contents of the resources and entities layers of the system. The *content processing pipeline* of figure 1 is then activated in order to populate the remaining mentions layer, to add missing entities and to establish the links among resources, mentions, entities and contexts. The remainder of the section describes the processing performed by the four main modules of the pipeline.

Content extraction. The first step of the content processing pipeline consists in extracting mentions and their attributes from stored multimedia resources, in order to populate the mention layer of the KNOWLEDGESTORE. Content extraction from texts and speech transcriptions is performed using the TextPro suite[6] (Pianta et al., 2008). TextPro supports multiple languages; for the Italian language, it has been trained on the Italian Content Annotation Bank (I-CAB) (Magnini et al., 2006), a resource containing 525 Italian pieces of news with manually annotated mentions. TextPro is used to locate and classify PER, ORG and GPE/LOC mentions. In addition, it has been configured to extract the `triggers` of PER mentions and to process them to extract mention attributes such as a person's `activity`, `nationality` and `title` (see figure 2). For example, given the text "young president Barack Obama", TextPro recognizes "Barack Obama" as a PER mention and "young president" as its `triggers`, from which attributes `ageCategory` and `activity` are respectively populated with "young" and "president".

Coreference. The next step consists in coreferring the mentions extracted in the previous phase according to the real-word entity which they refer to. That is for example telling whether the "motor racer Valentino Rossi" in one document is the same as the "Rossi" in another document. Basically, the task of coreferring mentions is a clustering

[6] http://textpro.fbk.eu/

task and the JQT2 system (Zanoli et al., 2011) has been used to cluster PER and ORG proper names. JQT2 is a Java implementation of the QT (Quality Threshold) algorithm (Heyer et al., 1999) and allows for two operations: re-clustering the whole dataset from scratch and incremental clustering which consists in updating clusters that are affected by the new data only.
In contrast to other systems which adopt a fixed similarity threshold to group mentions referring to the same person, the introduced system uses a threshold capable of changing its value on the basis of the ambiguity of the name as estimated by using external resources (i.e. phonebooks). For each name to be clustered the algorithm was provided with a specific threshold value and with a rich set of features (e.g. Named Entities, triggers) co-occurring with the name in the same document.
As regards the system annotation speed common values are about 500 mentions/sec whereas accuracy on person names is about 93%.
Disambiguation of GPE/LOC mentions has instead been done by using Geocoder (Buscaldi and Magnini, 2010), a system designed for the coreference of ambiguous toponyms (e.g. "Cambridge" in UK or USA or "Alabama" as a river or a state). Geocoder uses geometric methods on top of the GeoNames[7] geographical database and the Google Maps geo-referencing service[8]; as a side effect, it is often able to disambiguate mentions by linking them to well-known toponyms in GeoNames.

Mention-entity linking. In this step, each cluster of coreferring PER and ORG mentions is linked to the corresponding entity in the KNOWLEDGESTORE background knowledge, if any, and the associations are stored back in the system (relations `mention` and `entity` in figure 2). Linking of GPE/LOC mention clusters is not considered, as Geocoder already links them to GeoNames toponyms and we assume that GPE/LOC entities in the background knowledge, if any, are already aligned to the corresponding GeoNames toponyms (e.g., via `owl:sameAs` triples).
As multiple entities may have a name compatible with the surface form of a mention, the problem of *ambiguity* arises. This is addressed using a *context-driven* entity linking algorithm (Tamilin et al.,) that processes each mention cluster in two step. First, the KNOWLEDGESTORE contexts that more closely match the textual contexts of mentions are identified and the associations are stored in the system (relation `context` in figure 2); matching is performed based on the `subject`, `time` and `location` values automatically extracted from resources and their metadata. Then, only selected contexts are searched for the matching entity, thus helping with disambiguation by reducing the number of candidates. Candidate entities are ranked based on the similarity of their description with the surface forms, the metadata and the textual context of mentions: above a certain confidence threshold, the best candidate is chosen as the target of linking.

Entity Creation and Naming. For each unlinked mention cluster, a new entity is created and stored, with the goal of populating the KNOWLEDGESTORE with descriptions of real-World entities whose existence is unknown in the background knowledge but can be inferred from stored resources. A triple is stored to denote the name of the new entity, which is chosen based on the surface forms of the mentions in the cluster: longer and frequently mentioned names are preferred. As a future development, we plan to aggregate the attributes stored at mention-level to extract additional triples, both for new and already defined entities.

3. The Trentino KNOWLEDGESTORE

In the scope of the *LiveMemories* project, an instance of the KNOWLEDGESTORE system has been created and populated with multimedia contents and manually-acquired background knowledge relevant to the Italian Trentino region; the language for textual and audio contents is Italian. In this section we report on this experience, covering with detailed statistics and a running example the initial population of the instance (section 3.1) and the application of the modules forming the content processing pipeline (sections 3.2 – 3.5).

3.1. Initial Population

The KNOWLEDGESTORE instance has been initially loaded with 100K multimedia resources and with ∼350K triples of background knowledge describing ∼30K entities.
Multimedia resources consist of written news and images from three daily and weekly local newspapers – l'Adige[9], VitaTrentina[10], Federazione Trentina della Cooperazione (Fed. Coop.)[11] – as well as videos of daily television news from the local television RTTR[12]. Statistics about loaded contents are reported in table 1.

Table 1: Resources statistics.

Provider	News	Images	Captions	Videos
l'Adige	733,738	21,525	21,327	-
VitaTrentina	33,403	14,198	7,516	-
RTTR	2,455 (*)	-	-	120 h
Fed. Coop.	1,402	-	-	-
Total	770,998	35,723	28,843	120

(*) As result of the transcription segmentation.

Background knowledge has been manually collected from several Web sources, including the Italian Wikipedia, sport-related community sites and the official Web sites of local and national-level Public Administrations, economic and government bodies (e.g. the Italian Parliament Web site). Table 2 reports the number of loaded contexts, entities and triples aggregated along top-level topics.
Figure 3 introduces our running example and shows how resources are managed in the KNOWLEDGESTORE. The image shows the media objects providing the metadata for a video from RTTR (media R1), that has been automatically transcribed (media R2) and a news (media R3) containing an image (media R4) and its caption (media R5)

[7] http://www.geonames.org/
[8] http://code.google.com/apis/maps/
[9] http://www.ladige.it/
[10] http://www.vitatrentina.it/
[11] http://www.ftcoop.it
[12] http://www.rttr.it/

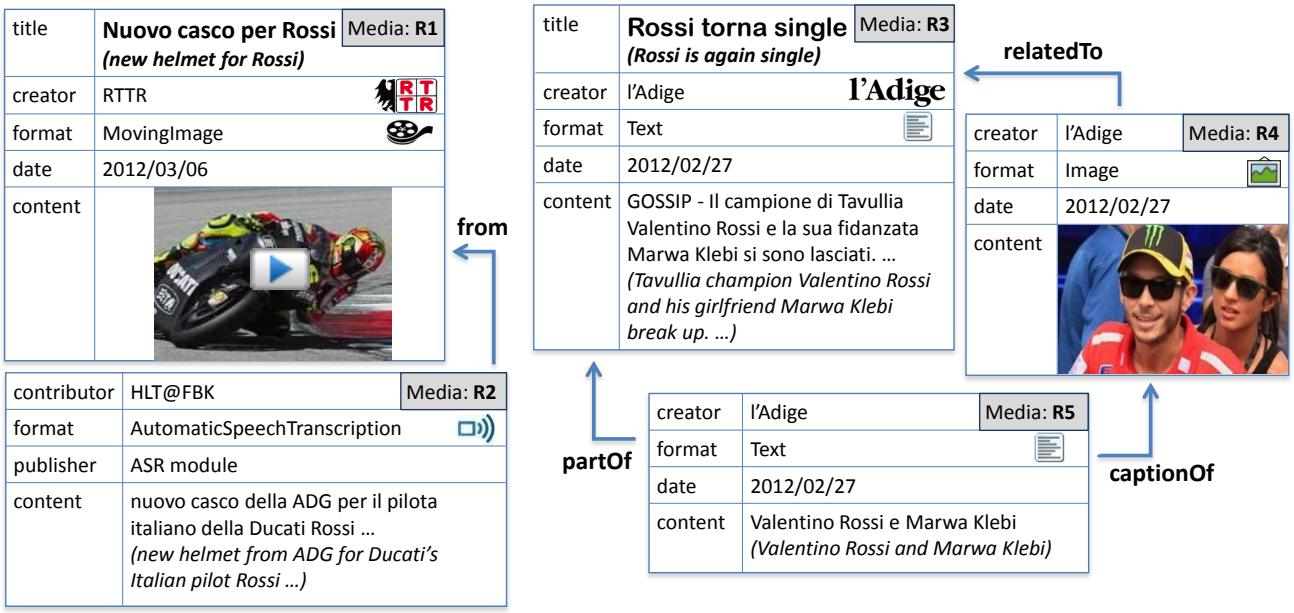

Figure 3: Running example – Examples of resources and their media content.

Table 2: Background knowledge statistics.

Main topic	Contexts	PER entities	ORG entities	(1) Avg. preds.	(2) Total triples
sport	136	8,570	191	3.81	192,115
culture	20	9,785	1	2.00	33,236
justice	7	354	10	2.16	1,575
economy	7	49	1,203	4.47	11,147
education	6	850	82	2.35	3,573
politics	535	8,402	319	4.64	98,780
religion	3	1,391	0	1.67	12,855
Total(3)	714	28,687	1,806	3.64	352,244

(1) Avg. number of distinct predicates for each entity.
(2) A triple mentioned in multiple contexts is counted once.
(3) An entity mentioned in multiple contexts is counted once.

Table 3: Content Extraction statistics.

Provider	PER	ORG	GPE+LOC
l'Adige	5,387,994	3,100,994	3,052,011
VitaTrentina	144,486	100,789	136,611
RTTR	19,290	15,493	27,404
Fed. Coop.	14,404	12,731	8,513
Total	5,566,174	3,230,007	3,224,539

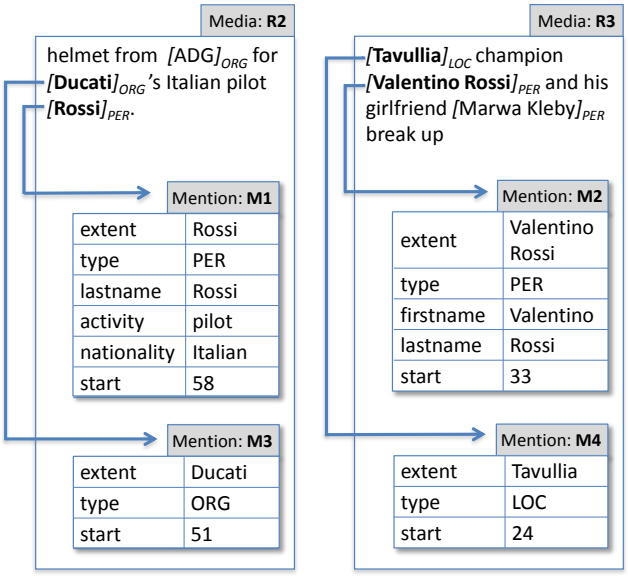

Figure 4: Running example – Examples of mentions as they are extracted and stored in the KNOWLEDGESTORE.

from l'Adige. Furthermore all the relations among medias are specified: the transcription is derived from the video, the caption is associated to the related image and it is part of the news, and the image is related to the news.

3.2. Content Extraction

In the text media that we considered (see table 1), around 10% of the words are part of PER, ORG and GPE/LOC mentions. These mentions have been extracted using TextPro and table 3 provides the number of retrieved mentions by data provider and entity type.

Continuing the running example, figure 4 shows how the transcription of media R2 and the text of media R3 have been annotated using TextPro (mention are enclosed in brackets). Starting from these two medias, the four mentions M1-M4 depicted in the figure are extracted and stored in the KNOWLEDGESTORE, enriched with semantic information specified by nearby trigger words.

3.3. Coreference

Starting from over 12 millions of recognized mentions in texts, we used the cross-document coreference system described in Section 2.2 to recognize more than 400,000 different clusters. Figure 5 reports the output of the system for

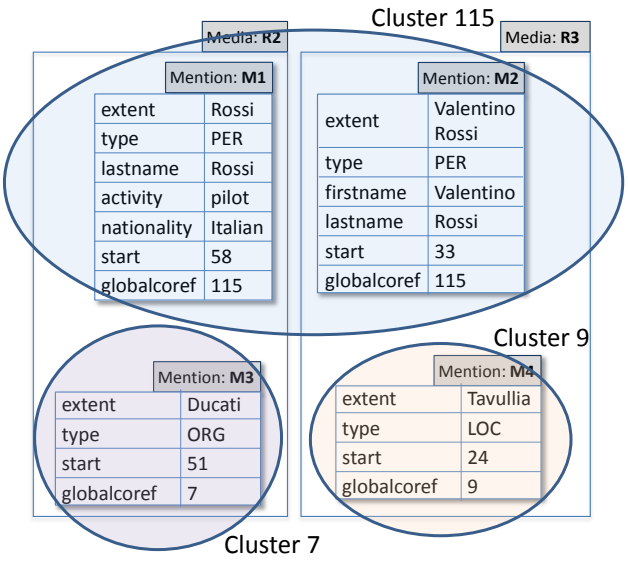

Figure 5: Running example – Cross-document coreference. Mentions "Rossi" (M1) and "Valentino Rossi" (M2) have been put in the same cluster and assigned with the same `globalcoref` identifier 115. In contrast "Ducati" (M3) and "Tavullia" (M4) set up the new clusters 7 and 9.

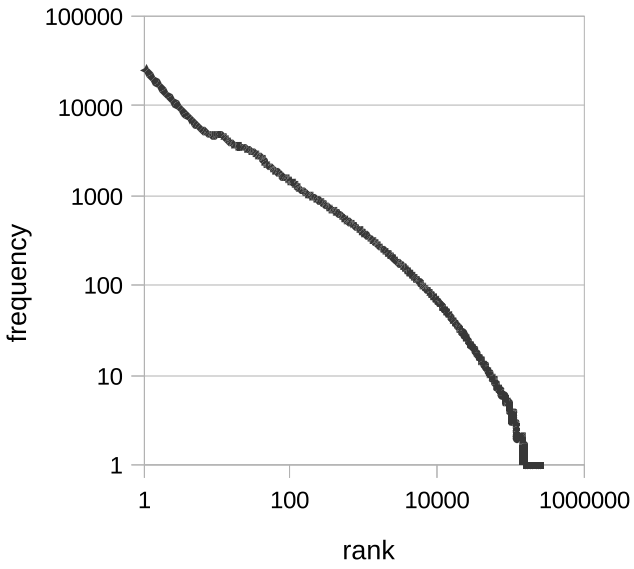

Figure 6: Persons Zipf's law. The graph shows rank versus frequency using logarithmic scales.

the running example: the two mentions "Rossi" (M1) and "Valentino Rossi" (M2) have been grouped together (they are thought to be referring to the same entity), whereas "Ducati" (M3) and "Tavullia" (M4) form distinct clusters.
Table 4 shows the number of different PER, ORG and GPE/LOC clusters as result of the coreference phase.
Concerning GPE/LOC mentions, it is worth noticing that 48.64% of clusters produced by Geocoder have been also linked to GeoNames toponyms, thus disambiguating them (in terms of mentions, the percentage increases to 65.04%).
As far as the PER entity names are concerned, their distribution in texts has been studied ranking them according to their frequencies and listing the frequencies in descending order. Figure 6 shows that the distribution follows the Zipf's law.

Table 4: Coreference statistics.

	PER	ORG	GPE+LOC	Total
Mentions	5,566,174	3,230,007	3,224,539	12,020,720
Clusters	340,147	16,649	52,478	409,274

3.4. Linking

The linking of PER and ORG mention clusters to entities and contexts in the background knowledge is exemplified in figure 7. Cluster 115 with mentions M1 and M2 in the running example is linked to entity E1 for pilot Valentino Rossi. The link is stored in the KNOWLEDGESTORE by associating both cluster mentions to E1. As part of the linking process, mentions M1 and M2 are also associated, respectively, to contexts C1 and C4 that more closely match their textual contexts: this information can be used, for instance, to extract and display only triple ⟨ `E1`, `team`, `Ducati`⟩ when presenting the media with mention M1 to a user.
Table 5 reports the percentage of clusters and mentions linked by the system – i.e. the *linking coverage* – for the different types of entities. Linking coverage is low in terms of entities and increases in terms of mentions, indicating that only the most popular (and thus frequently mentioned) entities are present in the background knowledge.
As the KNOWLEDGESTORE instance contains a super-set of the news used in the Evalita 2011 evaluation campaign[13], a rough evaluation of the quality of discovered links for PER mentions has been performed against a gold standard derived from the one used in Evalita for cross-document coreference. The gold standard built consists of 21,273 textual news from L'Adige containing 22,511 PER mentions grouped in 298 clusters, 73 of which have been manually associated to corresponding entities in the background knowledge. The linking system scored 84.5% accuracy (i.e. the fraction of correct link/not-link decisions), 85.1% precision (i.e. the fraction of linked clusters whose target entity is correct) and 89.5% recall (i.e. the fraction of manually established links discovered by the system).

Table 5: Linking coverage.

	PER	ORG	Total
Linked clusters	5.03%	7.96%	5.17%
Linked mentions	22.36%	12.02%	18.58%

3.5. Entity Creation and Naming

As the final step of the pipeline, starting from unlinked mention clusters new entities have been added to the KNOWLEDGESTORE instance. For example, unlinked cluster for mention M_4 in the running example is mapped to a new entity E_2, whose name "Tavullia" is chosen based on the mention surface form and stored using a `name` triple.
Table 6 presents the situation of the KNOWLEDGESTORE after the addition of new entities, reporting the numbers

[13] http://www.evalita.it/2011

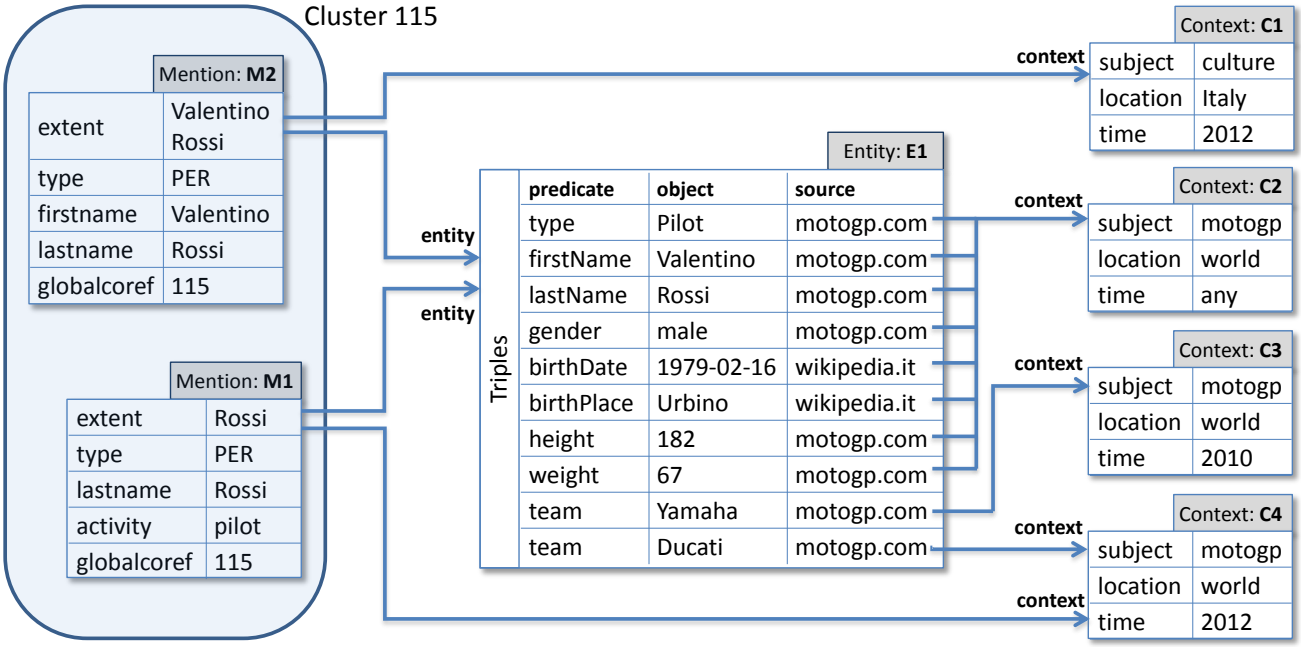

Figure 7: Running example – Linking to background knowledge.

of stored entities by type and provenance (media vs background knowledge, the latter either linked or unlinked to mention clusters). Entities induced from media correspond to 92.76% of all stored entities. The large percentage highlights the limits of manually acquired background knowledge, which can only cover the most popular entities. It also suggests that there is a large potential for applying knowledge base population techniques on stored multimedia resources, in order to (semi-)automatically populate the remaining long tail of less popular entities.

Table 6: Entities statistics.

Entity provenance		PER	ORG	GPE+LOC	all types
BK(*)	unlinked	11,566	480	–	12,046
	linked	17,121	1,326	–	18,447
media induced		323,026	15,323	52,478	390,827
all provenances		351,713	17,129	52,478	421,320

(*) Background knowledge.

4. Related Works

The attempt to build a pipeline for ontology population tasks in not new in the Natural Language Processing community. On the Content Information Extraction side, although several linguistic taggers are available for a number of languages (e.g. OpenCalais[14], Gate[15]), still there are no attempts to systematically store and make available the annotated data on a large-scale, and to integrate such data with background knowledge through Semantic Web technologies. As a matter of fact, much more attention has been paid on the processing side (e.g. several taggers are offered as web services) rather than on the storage side. The focus has been mainly on formats for single tasks, without a clear overall design, with the consequence that the interaction between linguistic, semantic, and world knowledge is still underspecified and poorly investigated. In this direction, the KNOWLEDGESTORE is intended to exploit the benefits of a common place for representing linguistic, semantic, and world knowledge on a large-scale.

On the architecture side, a popular annotation framework is UIMA[16] – Unstructured Information Management Architecture (Ferrucci and Lally, 2004) – developed by IBM and used as architectural basis for several NLP systems, including the recent Watson system. UIMA enables applications to be decomposed into components, and each of them implements interfaces defined by the framework and provides self-describing metadata via XML descriptor files. UIMA might be considered as complementary to the KNOWLEDGESTORE in that it provides a general metadata schema (called CAS) which might eventually be adopted by our system, although there is no specific reference to the representation of large amounts of knowledge.

A recent initiative which shares some of the motivations with the KNOWLEDGESTORE is the NLP Interchange Format (NIF)[17]. NIF is an RDF/OWL-based format developed by the University of Leipzig that aims to achieve interoperability between Natural Language Processing tools, language resources and annotations. The core of NIF consists of a vocabulary, which can represent strings as RDF resources. A special URI Design is used to pinpoint annotations to a part of a document. These URIs can then be used to attach arbitrary annotations to the respective character sequence. Employing these URIs, annotations can be published on the Web as Linked Data and interchanged between different NLP tools and applications. Although NIF

[14] http://www.opencalais.com/
[15] http://gate.ac.uk/
[16] http://uima.apache.org/index.html
[17] http://nlp2rdf.org/nif-1-0

is very recent (November 2011) and there is no enough feedback from the scientific community, the main motivations are shared by the KNOWLEDGESTORE approach we have presented in this paper.

As for the Semantic Web community and the Linked Open Data initiative, a significant research effort has gone along the direction of supporting the interlinking of knowledge and text documents at the representation and publication levels. This includes the standardization of metadata ontologies for describing generic information resources (e.g. Dublin Core) and multimedia contents (Suãrez Figueroa et al., 2011), and the deployment of mechanisms such as GRDDL (Connolly, 2007) and RDFa (Pemberton et al., 2008) for embedding RDF statements in XML and HTML documents, and HTTP content negotiation for relating the RDF description of an entity with its corresponding human-readable unstructured representation. All these solutions are complementary to the KNOWLEDGESTORE and can be used to expose its contents both for human and machine consumption, achieving interoperability with Semantic Web applications.

Aiming at bridging the gap between the Web of documents and the Web of Data, several Web services have been developed for recognizing mentions of named entities in an input text and linking them to URIs in relevant Linked Data datasets, such as DBPedia Spotlight[18] (Mendes et al., 2011), Zemanta[19] and AlchemyAPI[20]. Most of these services rely on the direct or indirect link between Web of Data entities and corresponding Wikipedia pages for disambiguation, whereas the KNOWLEDGESTORE linking module relies on the contextualization of knowledge.

Finally, it is worth noting that also in the Semantic Web community little attention has been paid to the storage of semantic data interlinked with multimedia resources. While *triple stores* have evolved into scalable solutions for storing, querying and reasoning with large amounts of knowledge, they currently provide only a limited support for integrating knowledge with multimedia, often consisting in simple full text search capabilities on RDF literals.

5. Conclusions

With its management of fine-grain links between knowledge and multimedia resources, the KNOWLEDGESTORE represents a valuable resource we plan to exploit to further investigate the mutual benefits of integrating multimedia and knowledge processing. In particular, three relevant research directions supported by the KNOWLEDGESTORE are (1) the development of *semantic search* approaches to jointly access multimedia and knowledge contents, (2) the 'injection' of background knowledge in NLP tools to improve their performances, as proven by recent works on coreference resolution (Bryl et al., 2010) and (3) the development of *knowledge base population* methods that build on the existing links between knowledge and multimedia to further enrich stored knowledge with information extracted from texts and other multimedia contents. Moreover, we plan to realize a feature outlined in the requirements for the KNOWLEDGESTORE (cfr. section 1) but not implemented yet, namely the ability to populate the KNOWLEDGESTORE *incrementally*.

Concerning the developed software, we intend to release the code of the KNOWLEDGESTORE *core* as an open-source distribution.

[18] http://dbpedia.org/spotlight

[19] http://www.zemanta.com

[20] http://www.alchemyapi.com

6. Acknowledgements

This work was partially supported by the LiveMemories[21] project (Active Digital Memories of Collective Life) funded by Autonomous Province of Trento (Italy) under the call "Major Projects".

[21] www.livememories.org

7. References

V. Bryl, C. Giuliano, L. Serafini, and K. Tymoshenko. 2010. Using background knowledge to support coreference resolution. In *ECAI'10*, pages 759–764.

D. Buscaldi and B. Magnini. 2010. Grounding toponyms in an italian local news corpus. In *GIR'10*.

J. J. Carroll and G. Klyne. 2004. Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation.

D. Connolly. 2007. Gleaning resource descriptions from dialects of languages (GRDDL). W3C recommendation.

D. Ferrucci and A. Lally. 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348.

T. Heath and C. Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool Pub.

L. J. Heyer, S. Kruglyak, and S. Yooseph. 1999. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 9(11):1106–1115.

B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, Bartalesi Lenzi, and R. Sprugnoli. 2006. I-cab: the italian content annotation bank. In *LREC'06*.

P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *I-Semantics'11*, pages 1–8. ACM.

S. Pemberton, B. Adida, S. McCarron, and M. Birbeck. 2008. RDFa in XHTML: Syntax and processing. W3C recommendation.

E. Pianta, C. Girardi, and R. Zanoli. 2008. The TextPro tool suite. In *LREC'08*.

L. Serafini and M. Homola. 2011. Contextual representation and reasoning with description logics. In *24th Int. Workshop on Description Logics (DL 2011)*.

M. C. Suãrez Figueroa, G. A. Atemezing, and O. Corcho. 2011. The landscape of multimedia ontologies in the last decade. *Multimedia Tools and Applications*, 55(3), 12.

A. Tamilin, B. Magnini, and L. Serafini. Leveraging entity linking by contextualized background knowledge: A case study for news domain in italian. In *6th Workshop on SW Applications and Perspectives (SWAP)*.

R. Zanoli, F. Corcoglioniti, and C. Girardi. 2011. Dynamic threshold for clustering person names. In *Proceedings of EVALITA 2011*.