LIVEMEMORIES

Active Digital Memories of Collective Life

WWW.LIVEMEMORIES.ORG

MAJOR PROJECTS 2006
DEL. OF "GIUNTA PROVINCIALE DI TRENTO" NR. 686 OF MARCH 18, 2008

Coordinator: Fondazione Bruno Kessler (FBK)
Via Santa Croce, 77 - 38100, Trento (Italy)
Scientific Coordinator: Bernardo Magnini

Partners: University of Trento (UNITN)
Via Belenzani, 12 - 38100, Trento (Italy)
Coordinator: Massimo Poesio

University of Southampton (SOTON)
University Road, Southampton SO17 1BJ (UK)
Coordinator: Wendy Hall

Contact: Bernardo Magnini (magnini@fbk.eu)

Start Date: October 1st, 2008
Duration: 36 months

Summary

1. Introduction

In the digital age, our records of past and present are growing at an unprecedented pace. Huge efforts are under way in order to digitize data now on analogical support; at the same time, low-cost devices to create records in the form of e.g. images, videos, and text are now widespread, such as digital cameras or mobile phones. This wealth of data, when combined with new technologies for sharing data through platforms such as Flickr, Facebook, or the blogs, open up completely new, huge opportunities of access to memory and of communal participation to its experience.
This huge amount of data are however unconnected, scattered across unrelated sites, often incomplete and inconsistent, and realized with different media. There is a need for ways to make these unrelated digital fragments to "talk" to each other. The scientific challenge for LiveMemories will be to develop methods for automatically interpreting the content of such fragments, and transforming into "active memories" this huge amount of multimedia data, pieces that, as in an immense digital puzzle of the collective memory, are to be integrated, completed and adapted, in order to rebuild the pictures of our past. As an example, consider a portal set up by a school for students to create a story of that institution by sharing their texts, photos and videos. A student might a picture with a tag like "Mario Rossi with his fellow students in 3A at the Da Vinci school in 1972". In a 'live' memory, photograph and caption would be turned from static information into active content by interpreting text and image, and linking them to other memories in the portal—for instance, to the records of 3A in 1972, recent pictures of these student, perhaps some information about them from the papers, etc.
The challenge for LiveMemories will be the nature of the data which constitute digital memories: data which

- come from heterogeneous sources (e.g. written texts, radio news, websites, images and videos, maps, documents in different languages);
- are dynamic, i.e. subject to change in time, and therefore need constant update and integration with new knowledge;
- are often inconsistent or incomplete, which mines those models of knowledge which are too rigid;
- have to be adapted to users' views.

In addressing these problems, the project will take advantage of the outstanding competence in Human Language Technologies, Knowledge Management, and Web Science of the partners. We specifically aim at improving the state of the art in the following three areas: content extraction from different multimedia sources, content integration through the application of large-scale reasoning techniques, content presentation through the realization of adaptive views. A concrete and permanent result will be the Content Management Platform, through which it will be possible to daily monitor and store contents from a number of Trentino-based media (a showcase in collaboration with the Trentino Technological District is scheduled).
The effectiveness and social impact of LiveMemories will be assessed through a variety of studies of how collective memories are gathered. A local community will be involved in a study. A web portal will be made available to the general public for the collection,

management, integration and fruition of multimedia collective memories coming from different sources (e.g. associations existing on the territory, local newspapers and radios, etc.), involving also families and single persons in the objective of making fragments of collective history available. Visibility will be achieved through these initiatives and through a LiveMemory exhibition for the general public, set up to show that the technological advancements underlying the idea of memories as pieces of active digital content open the opportunity of new modalities of experiencing collective memories of our past with a high social impact.

2. General objective

From a scientific / technical perspective, LiveMemories aims at scaling up content extraction techniques towards very large scale extraction from multimedia sources, setting the scene for a Content Management Platform for Trentino; using this information to support new ways of linking, summarizing and classifying data in a new generation of digital memories which are `alive' and user-centered; and to turn the creation of such memories into a communal web activity. Achieving these objectives will make Trento a key player in the new Web Science Initiative, digital memories, and Web 2.0, thanks also to the involvement of Southampton. But LiveMemories is also intended to have a social and cultural impact besides the scientific one: through the collection, analysis and preservation of digital memories of Trentino; by facilitating and encouraging the preservation of such community memories; and the fostering of new forms of community, and enrichment of our cultural and social heritage.

3. State of the art

We summarize state of art in four areas: Digital Memories, the main application area, and three areas of scientific activity: multimedia content extraction, content integration and content presentation.

DIGITAL AND COMMUNAL MEMORIES
The digital revolution creates new opportunities for collecting and displaying memories of the past. Digital memories may make certain heritage or historical artifacts (e.g., books too old to be handled) accessible to the non-specialist, and offer professionals reduced search times. This promise is recognized by the EU Commission, which has made digital libraries a key aspect of the i2010 initiative. Among the many current initiatives in the area we will mention:

- MEMORIA PER IL TRENTINO (www.trentinocultura.net/memoria/hp_memoria.asp) a project devoted to the discovery, protection and promotion of historic, artistic and cultural heritage, especially in relation to the last century, which has seen the local community maintain its own identity in spite of profound changes

- MALACH (malach.umiacs.umd.edu/) in which new technologies in automatic speech recognition and computer-assisted translation have been developed to increase access to archives of videotaped oral histories assembled by the Shoah Visual History Foundation.
- MEMORIES FOR LIFE (www.memoriesforlife.org/), which investigates what human and computerized memory have in common so as to develop technologies to store and retrieve memories more efficiently across personal, social and work domains.
- E-MAIL BRITAIN, started in May 2007, a British Library initiative to collect 1M emails on various topics from the general public.

The opportunities afforded by digital forms of memory preservation are even more promising in the move to web 2.0— web-based platforms to share and communally create information, such as Wikipedia, Facebook, or Flickr. Public participation in memory creation may result in unearthing valuable documents not available to institutions.
The interest in digital memories and communal memories is highlighted by the number of events on the theme, such as the CARPE workshops which cover capture, retrieval, organization, search, and privacy and legal issues related to the continuous archival and retrieval of all media relating to personal experiences, or CIRN (Community Informatics Research Networks) (October 2006), on the theme "Constructing and sharing memory: Community informatics, identity and empowerment", focusing on how we construct memory, what is the role of community informatics in the development of new means to capture private and public memory, and how information and communications technologies assist communities to use memory.

CONTENT EXTRACTION
LiveMemories will extract digital content from three sources: text, automatic speech transcriptions, images.

Content extraction from text
The availability of high-performance tools for POS tagging and parsing has made it possible to contemplate large-scale semantic processing (named entity extraction, coreference, relation extraction, ontology population). US initiatives such as MUC, ACE and GALE made large annotated resources available and introduced quantitative evaluation. In intra-document coreference (IDC) this led to the development of the first large-scale machine learning models using these resources [9.12, 9.13] and to the development of IDC tools—most recently, the ELKFED/BART system (www.clsp.jhu.edu/ws2007/groups/elerfed/). In relation extraction, work carried out as part of the ACE initiative and in ELERFED showed that good results can be obtained extracting relations from news with supervised methods, particularly Support Vector Machines (SVMs) and Kernel Methods [10.1] but that semi-supervised methods are more effective with less formal text.
The performance and usefulness of existing tools can be improved by:

- Larger corpora and techniques obviating the need for large scale annotation (e.g., active learning, weakly supervised methods).
- Better preprocessing techniques (often underestimated);
- Incorporating automatically extracted lexical and commonsense features in addition to traditional ‘surface’ features;

- Developing better Machine Learning methods to exploit these more advanced sources of information (e.g. kernel functions).
- Developing richer representations of relations, e.g., with temporal modification (e.g. John Doe was CFO of ACME from 2001 to 2005);
- Further developing automatic methods for Textual Entailment Recognition [9.12], a robust type of textual inference based on patterns that can be automatically acquired from corpora.

Content extraction from speech
Speech recognition has advanced to very large vocabulary, speaker-independent connected-speech recognition. While public awareness of the technology remains low, text dictation in professional settings is now often supported by speech recognition. The research community has recently focused on speakers that do not speak with the intention of being recognized by an automatic system, e.g. news, conversations, speeches. Typical application scenarios meeting transcriptions or audiovisual digital library indexing.
Speech Analysts investigate the challenge of mining speech transcriptions to extract relevant information.
Open issues are that sentence boundaries, capitalizations, and intonations are to be automatically reconstructed before text processing.

Content extraction from images
In many cases, special-purpose image processing systems produce high performance, particularly when simplifying assumptions can be made about the domain. To take three key examples:

- Very simple patterns can be created alongside the systems to process them, such as automatic barcone readers.
- Face recognition techniques (and other visual biometric systems) are reliable enough for widespread commercial use.
- Medical diagnosis from images (e.g. mammograms) is another important area for study.

The LiveMemories domain makes strong demands in terms of the breadth of the envisaged application: content is likely to be diverse, there is a variety of formats, languages, tagging vocabularies and so on.
The exploitation of ontologies and other technologies associated with the Semantic Web is important for this research problem. Browsing and indexing techniques, together with methods to convert metadata into well-structured RDF, can be used to support structured querying of multimedia databases.
Furthermore, a large amount of data is available in digital images via automatically-gathered metadata using formats such as the Exchangeable Image File Format (Exif); e.g., the GPS location of the image, the time and focus information. The combination of straightforward feature detectors (e.g. edge detectors) with associated information (e.g. online news or weather reports), shows that the scope for intelligent guesswork about image content is wide.

CONTENT INTEGRATION
Content integration involves the ability to automatically recognize coreference among entities, to assign appropriate geographical and temporal tagging to pieces of content, to

represent and store such content in ontologies, and to retrieve it taking advantage of reasoning and inferences.

Cross-document Coreference

Interest in cross-document coreference has began fairly recently [9.14], but there has been much development in recent years because of great interest both from government and from industry. In particular there has been great interest in a simpler form of entity disambiguation, generally known as Web entity as in the case of the Web people task of Semeval [9.15] and the Spock challenge (challenge.spock.com/). As testified by the SEMEVAL Web People task, most state of the art systems are based on clustering of entity descriptions containing a mixture of collocational and other information, among which information about entities and relations. SEMEVAL also showed that the clustering technique and especially the termination criterion are crucial. Finally, work on the Spock challenge highlighted the need for methods for handling huge quantities of information. Recent developments have therefore focused on improving the clustering technique and experimenting with different types of information that can be extracted robustly from text. (See, eg., the results with ELERFED.) Most of the work discussed above was carried out for English; progress with languages other than English includes work on German (e.g., Versley) and Spanish (Ferrandez) but very little on Italian apart from work by Delmonte [9.16], also in part for lack of resources.

Complex reasoning services

In Knowledge Representation and Reasoning (KRR) there are a number of mature formalisms and reasoning tools that are currently used also in real world applications. Currently, Description Logics [9.7] is the most diffused formalism to represent ontological knowledge. Efficient reasoning tools like Pellet, RacerPro, Fact++, have been developed and made available for most of the applications in the semantic web. Scalability of this approach has been reached through modularity, i.e., knowledge is organized in relatively small interconnected modules. Drago [9.8], developed at FBK, is a tool that support reasoning in modular ontologies. Description Logics has been extended to deal with uncertain knowledge [9.9,9.10]. In the representation of time dependent knowledge, temporal databases provide an underlying well established technology, which, in principle, is relevant to the LiveMemories applications. More recently, there has been work on OWL-MeT (ermolayev.com/owl-met/) an extension of OWL aiming at presentation of both topological and metric properties of time. Similarly extensions of the OWL language have been proposed to represent uncertain knowledge. An example in this direction is [9.9]. Representing propositional attitudes like belief, intentions, etc. has been the object of study for many years in KRR.

However, there is no mature reasoning tool for dealing with ontologies and propositional attitudes. Finally a language that allows to combine all the above reasoning services is still missing.

Robust Ontology Population

Ontology learning and population is one of the most emerging research theme in the semantic web.

Ontology learning has the objective of (semi) automatic construction of ontologies starting from semi-structured data. On the other side ontology population has the main objective of finding both individuals relations among them and then to populate the ontology. Text-to-Onto, Web->KB are two examples of learning tools based on Formal Concept Analysis and Bayesian learning methods (see [9.10] for a tutorial on ontology learning from text).

Experience of ontology construction and population has been carried on at FBK withing the Ontotext project. See for instance [9.11]

CONTENT PRESENTATION
Existing techniques for allowing access to large diverse multi-dimensional content collections include:

- Hierarchical classifications (HC): rooted trees in which nodes are assigned natural language labels (categories). The category of a child node represents either a sub-category or a specification of the parent's category. HC are a very natural way to organize and access content, and they are used in libraries, web search, e-commerce, personal knowledge management etc.
- Clustering [9.1]: a technique based on similarity measures for automatically grouping items. An example of clustering of Web results can be found at Clusty.com.
- Faceted classification [9.1]: allows describing a collection by means of several orthogonal characteristics common to all the items. Users can generate selective views by applying the characteristics in an arbitrary user-defined order.
- Navigation: allows discovering content categories that are related to a given content category, but not directly connected to it. Navigation can be applied within same HC or across multiple HC. It allows jumping from one node in a branch to node in a different branch, like e.g. the @-links in the dmoz open directory project [9.4].
- Syntactic search (SyS): an approach to finding content items by matching keywords and/or attribute values, provided by the user, with the body and/or attributes of the items in a collection. In SyS, matching is resolved by computing string relations.
- Semantic search (SeS) refers to the use of semantics in query construction, in the search process, and in the visualization of search results [9.2]. In query construction, semantics is enabled through the use of controlled vocabularies, query disambiguation, and by applying semantic constraints on the meaning of query terms. During search, semantics is implemented as query expansion, graph traversal, spread activation, RDFS/OWL reasoning. Despite about 35 SeS systems presented in the literature, there is evident lack of evaluation of semantic search algorithms, of user evaluations of the interfaces, and of APIs and middleware support [9.3].
- Social centered techniques: in recent years the Web 2.0 has revived community approaches previously suggested (in part) by CSCW. Among them we mention social filtering and recommendation systems, folksonomies, tagging and social bookmarking, social navigation. Novel approaches like "human computation" [9.5] are emerging in this area.

REFERENCES:

[9.1] Hearst, Clustering versus Faceted Categories for Information Exploration, Communications ACM 49(4), 2006
[9.2] Hildebrand et al., An analysis of search-based user interaction on the Semantic Web. Information Systems Center, INS-E0706, 2007
[9.3] Shvaiko et al., A Survey of Schema-based Matching Approaches. JoDS, IV, 146, 2005
[9.4] Glover et al., Using Web Structure for Classifying and Describing Web Pages, Proceedings of WWW-02, 2002

[9.5] van Han et al., Games with a purpose. IEEE Computer, 2006
[9.6] Giunchiglia et al., Semantic Matching: Algorithms and Implementation. JoDS IX, 2007
[9.7] Baader et al., The Description Logic Handbook. CUP, 2003
[9.8] Serafini et al., DRAGO: Distributed reasoning architecture for the semantic web. ESWC 05, 2005
[9.9] Straccia, Reasoning with Fuzzy Description Logics. JAIR14, 2001.
[9.10] Buitelaar, Ontology learning from text. Tutorial at ECML/PKDD, 2005
[9.11] Popescu et al., From Mention to Ontology: A Pilot Study. SWAP, 2006.
[9.12] Szpektor et al., Scaling Web-based Acquisition of Entailment Relations. EMNLP, 2004.
[9.13] Dagan et al., Direct Word Sense Matching for Lexical Substitution. ACL, 2006
[9.14] Bagga and Baldwin, Entity-based Cross-document Coreferencing Using the Vector Space Model. ACL, 1998
[9.15] Popescu and Magnini, IRST-BP: Web People Search Using Name Entities. SEMEVAL, 2007
[9.16] Delmonte et al., VENSES - a Linguistically-Based System for Semantic Evaluation. PASCAL, 2005
[10.1] Moschitti et al, Tree Kernels for Semantic Role Labeling. CL, 2008.
[10.2] Moschitti et al, Exploiting Syntactic and Shallow Semantic Kernels for Question/Answer Classification. ACL, 2007
[10.3] Moschitti and Zanzotto, Fast and Effective Kernels for Relational Learning from Texts. ICML, 2007
[10.4] Riccardi et al., NEEDLE: Next Generation Digital Libraries, AISV, 2006
[10.5] Tuffield et al., Towards the Narrative Annotation of Personal Information and Gaming Environments. Hypertext, 2005
[10.6] Domingue et al., Supporting Ontology Driven Document Enrichment Within Communities of Practice. Knowledge Capture, 2001
[10.7] Moschitti et al., Spoken Language Understanding with Kernels for Syntactic/Semantic Structures. ASRU, 2007

4. Preliminary results

The proposed project builds on top of, and integrates, the results of substantial technological development by the participating partners, in a number of projects.

CONTENT EXTRACTION

Text Processing

On the data cleaning and corpus collection side, UniTN has been a key participant in the Wacky corpus creation campaign and the CLEANEVAL initiative, producing also Jboot, a tool for automatic collection of web-based corpora. I-CAB, the Italian annotation bank (FBK), is a collection of 500 news from the Adige newspaper semantically annotated at several levels, including entity mentions (both descriptions and named entities) and temporal expressions. UniTN has many years of experience with annotating coreference and relations for both Italian and English. FBK also realized MultiWordNet, the Italian version of the English WordNet.

Tools developed include TextPro, developed at FBK, a suite of tools for processing Italian that includes POS tagging, lemmatization, and named entities recognition for Italian; tools, based on the Lucene platform, for indexing and retrieving large-scale document collections. A solid foundation for further IDC work has been established as part of the already mentioned ELERFED workshop at Johns Hopkins University, funded by US NSF and US Department of Defense (UniTN), during which the BART tool was developed, currently being made ready for open source release via SourceForge. Tease is a tool for acquisition of entailment patterns from the Web (FBK). FBK has been the organizers of Evalita 2007 (evalita.itc.it), a comparative evaluation of Italian NLP tools.

Speech processing
FBK has a long tradition in the development of leading-edge technologies for automatic speech recognition. In particular, its technology for automatic transcription of speech found in audio streams has been recently refined and assessed within the EU project TC-STAR. This technology includes modules for automatic partitioning of the audio streams, decoding with large vocabulary, speaker/environment adaptation, and word graph re-scoring. For commercial exploitation, the developed technology is licensed to the FBK's spin-off PerVoice.

Image processing
Soton has many years' experience in the extraction of content from multimedia. Examples include: MAVIS, an architecture within which separate modules for processing associated with media-based feature types could be added and integrated; SCULPTEUR, which used Semantic Web technology for 3D museum object retrieval; MIAKT which used SW and knowledge technologies to support and manage collaborative medical image analysis; PHOTOCOPAIN, a semi-automatic image annotation system to combine information about context with other readily available information in order to generate outline annotations that the user may further extend or amend.

CONTENT INTEGRATION
During the ELERFED workshop cited above, several Entity Disambiguation systems were built in collaboration with Gideon Mann and colleagues from University of Massachusetts at Amherst, and tested on the Spock data. FBK has been involved in entity disambiguation research for a few years both as part of the OntoText project and by participating in SEMEVAL Web People Search in 2007, scoring at the second position.
Geo-time tagging. The Chronos system, developed at FBK, allows the recognition and normalization of temporal expressions for Italian and English. It has been evaluated at the Evalita initiative, scoring at the first position.
Ontology population. In the Ontotext PAT funded project FBK has developed a tool called ontology repository, that supports the management of background knowledge in the form of OWL ontologies, and provides search and reasoning functionalities exploited for ontology population. In the APOSDLE (EU funded) project that supports e-learning, FBK has developed a tool for community ontology specification via wiki, and automatic translation into an OWL ontology. The DRAGO reasoning system, developed at FBK, support the distributed reasoning between multiple integrated and heterogeneous ontologies.

CONTENT PRESENTATION
Sweb (http://www.dit.unitn.it/~knowdive) is a data and knowledge management system developed at the University of Trento. It allows its users to create hierarchical classifications, populate them with documents of various kinds, instantly search classifications and documents, create navigational links between classifications and use these links for browsing and distributed searching, share classifications, documents and links with other users in a fully controlled manner, and perform other operations. Sweb is a semantics-enabled system because it has a background knowledge base which encodes facts about a particular domain and this knowledge base is used to support the user in data and knowledge management tasks. For instance, it helps the user to generate navigational links, to perform semantic search, to disambiguate and explicitly define the meaning of terms that the user use in classification labels, in document attributes, in search queries,

and so on. The knowledge base is fully configurable and can be updated by the user depending on a concrete task.
SMatch [9.6] is a schema-based matching tool developed at the University of Trento. It takes two graph-like structures (e.g., classifications) and returns semantic relations (e.g., equivalence, subsumption) between the nodes of the graphs that correspond semantically to each other. The relations are determined by analyzing the meaning (concepts, not labels) which is codified in the elements and the structures of the schemas/ontologies. In particular, labels at nodes, written in natural language, are translated into propositional formulas which explicitly codify the intended meaning of the labels. This allows for encoding the matching problem as a propositional unsatisfiability problem, which can then be efficiently resolved using (sound and complete) state of the art propositional satisfiability deciders.
Although it has been primarily designed to solve the semantic heterogeneity problem in data/information integration, SMatch can be used in other applications, e.g., finding users with similar interests, discovering navigational links between users' classifications, and so on. Currently, SMatch is a component in the Sweb system.

5. Overview

The project is organized into five main activities:

1. Content Extraction,
2. Content Integration,
3. Content Presentation,
4. Content Management Platform, and
5. Showcase and dissemination.

The first three activities cover the main scientific interests of the project; their purpose is to deliver research results (e.g. top level publications, organization of relevant scientific events) and new solutions to be incorporated in the integrated platform. We expect of course these solutions to be applicable to other contexts besides Digital Memories, where content extraction and integration are crucial. In addition, such research activities are intended to create the opportunity for graduate training, by employing PhD students to carry out these projects. The fourth activity, Integrated Platform, is devoted to the integration of the software solutions developed in other activities into a well-engineered system. Finally, several showcases will be built on top of the platform. The project is articulated in three main cycles of activity of one year each. During the first cycle we will focus on processing institutional sources of data, to begin extracting the 'background' data that users of LiveMemories will have access to; a first platform will also be developed by integrating existing components. During the second cycle, we will complete the first advanced visualization functionalities and develop a first Example Showcase to illustrate the functionalities of the platform, and establish contact with a potential community. Finally, in the third cycle, we will realize a Pilot Showcase with this community.

1.*Content Extraction.* At present it is possible to extract from text information on a large scale both about entities (persons, locations, organizations) and about simple relations

between entities (e.g. the affiliation of a person to an organization), but with a precision which is not very satisfactory. We aim to significantly improve the state of the art in information extraction and to extend its scope to spoken and/or multilingual documents. The main challenge we foresee is developing methods that can achieve high precision out of very heterogeneous and potentially very noisy data; we expect the main focus of the work to lie in the areas of data cleaning, intra-document coreference, and relation extraction. Well established content extraction techniques for visual documents (images and video) will be adapted to the digital memory setting, in particular with respect to information processing, integration, and access requirements.

2.*Content Integration*. The objective of this activity is to develop novel techniques for integrating vast amounts of time-stamped and possibly inconsistent knowledge extracted from a variety of sources into the knowledge base that will be the core of LiveMemories. One goal will be to develop novel methods for cross-document coreference, able to handle vast amount of data and temporally stamped information.
Novel forms of representation of entities will also be required: the standards currently adopted by the Web community (e.g. RDF, OWL) do not support enough robustness for managing and integrating large amounts of data automatically extracted from unstructured sources. The challenge of this activity is to define and experiment with a number of well founded extensions to current formalisms to deal with relevant phenomena such as the role of the context where information is placed, the uncertainty, incompleteness, and time dependent information. A further research direction that we intend to purse is the convergence between logic-based inference approaches and the recent trend emerged in Computational Linguistics based on textual entailment.

3.*Content Presentation*. An important goal of the project is to allow the creation of communities that will extract information related to themes of their interest, organize it and present it to others (e.g. in virtual exhibitions). In this sense the project has a strong Web 2.0 connotation, enhanced by the ability to use semantics and to present the material with advanced visualization techniques, such as (semantic based) adaptive presentation, a variety of summarization techniques including, for instance, producing a filecard of relational information about inviduals, the production of 'narratives' displaying how an individual or a community evolved over time, and automatic translation into languages other than Italian. Some of these activities will draw on existing technologies, while in others advancements will be made possible by the research performed by the participating groups. Research on maintaining multiple views of knowledge will be key, as well as research on semantic matching and conversational interfaces. Southampton's expertise with multimedia visualization will be key. Lessons learned from the many social software tools that emerged over the last few years will be precious for shaping the functionalities offered to the user and the interface.

4.*Content Management Platform*. The technological core of the project is a multimedia digital library for the acquisition, management and integration of a large amount of active collective memories. This platform will support massive content extraction, while also providing users the functionality to create and access such memories, establishing links between all sorts of information. The result will consist of a catalogue of the geo/time referenced entities (i.e. persons, locations and organizations) which are present in the area selected for the showcase, and of a map of the relations existing between such entities. An entity-centric view will be adopted, where information and knowledge are accessed and searched for by using identifiers (i.e. URI) to retrieve entities; in turn, entity profiles are

built by gathering from diverse knowledge sources what the Web has to offer about these entities. The project builds on top of technologies developed by the Ontotext project (knowledge extraction from text), GIS platform for territory maps, web community and image sharing platforms.

5.*Showcase and dissemination*. In LiveMemories the focus is not so much on the life of single individuals, as on the events of the collective life of a community. The goal of the project is to provide communities with state-of-art technology enabling the creative construction of a collective memory. A web portal will be made available to the general public for the collection, management, integration and fruition of multimedia collective memories coming from different sources (e.g. associations existing on the territory, local newspapers and radios, etc.), involving also families and single persons in the objective of making fragments of collective history available. In a carefully controlled experiment, a local community will be supported in the creation of its own collective memory. LiveMemories events will be set up to engage the local population and to show that the technological advancements underlying the idea of memories as pieces of active digital content open the opportunity to new modalities of experiencing collective memories of our past with an high social impact.

6. Originality and relevance

We are confident that LiveMemories will make several original contributions to Human Language Technology, Knowledge Management and Web Science, making the project timely and highly relevant.
Research topics of high importance have been chosen, for which innovative approaches have been adopted, combining data driven methods with knowledge-based techniques. The partners are international leaders in intra- and cross-document coreference, kernel methods applied to relation extraction, textual entailment, statistical machine translation, ontology matching and contextual reasoning, and multimedia processing and visualization.
First of all, we feel that the time has come to test content extraction and content integration from non-structured information sources on a large scale. While the research community has focused so far on benchmarks of limited size (e.g. ACE - Automatic Content Extraction – corpora) the truly novel challenge of the project is to apply such methods in a real scenario. The opportunity, and the ambition, is to apply state of the art technologies to a territory as large as Trentino, covering half a million people living on a 6,200 square kilometres. To our knowledge, this will be the first experiment ever attempted to detect, monitor, and extract entities and facts on such a large scale. An original approach is envisaged for the overall process of ontology population, where already recognized facts are dynamically used to feed learning algorithms. We see this objective as part of a long term plan for building a durable platform for content managing of the Trentino territory.
Secondly, the type of application we envisage, Digital Community Memories, is highly innovative, yet one whose full potential has not yet been fully exploited. While there are several initiatives aiming at collecting digital memories for specific events as static repositories, the original perspective of LiveMemories is to automatically interpret such memories and then link them to each other. We see such “live memories” as having high

potential for original forms of fruition, including automatically reconstructing timelines for events, associating people, organizations or locations to events, retrieving facts associated to specific places. If we consider this potential as integrated into web-based technologies, now accessible by large portions of citizens, where people can spontaneously form communities, the social and the emotional
impact can dramatically change the way we think of our memories in the digital era.
LiveMemories will also result in an original and cross-disciplinary perspective on a number of important areas of investigation. The possibility of applying state of the art technologies in Computer Science on a large scale will make new sociological investigations based on statistical methods possible. The amount of data made available from content extraction will force the development of new theories for robust reasoning under uncertain data. Although this is not a new topic in Artificial Intelligence, the models developed so far have only been tested on a small scale, and very little is known about how these models will scale up when millions of data items have to be managed. We expect from this research one of the major theoretical contribution of the project.
The integration of data from institutional sources and from private citizens and families is another original aspect of the project. To our knowledge none of the Web-based platforms that are now available for sharing content affords this opportunity. Under this new perspective we expect that communities will move from the personal dimension (e.g. family, friends) to the collective dimension (e.g. a quarter, a school, a town). The long term challenge (beyond the scope of the present project) is to solicit both institutional data providers, associations and citizens to think of the collection and preservation of digital memories as a collective enterprise.

7. Links to national and international research program

- i2010 Digital Library Initiative. Digital Memories are emerging within the i2010 as an application area concerning new web based technologies for collecting, managing and making available large repositories of multimedia content.
- Web Science Initiative. Among the central topics (see recent speech at the US House of Representatives by Tim Berners-Lee, director of W3C) there are the development of advanced methods for knowledge integration and the study of the social implications of the Web. This will be a central theme in LiveMemories.
- Entity Disambiguation. This area is considered strategic by the US government, as shown by increased funding for initiatives such as the Summer 2007 ELERFED CLSP workshop on Entity Disambiguation, coordinated by the UniTN group and the forthcoming ACE cross-doc coreference evaluation.
- RTE. The visibility of the Recognizing Textual Entailment challenge is also shown by the fact that the next edition will be organized by NIST as part of the new Text Analysis Conference, under the coordination of FBK and Bar-Ilan.
- FIRB-Israel project “Intelligent Technologies for Cultural Tourism and Mobile Education”. Live Memories will also coordinate with this project, involving FBK, UniTN, U. of Haifa and Bar Ilan, where the UniTN group applies ontology matching methods to museum interfaces, whereas FBK works on textual entailment.

- OKKAM. A newly funded, EU-IP OKKAM, coordinated by UNITN, aiming and implementing a worldwide infrastructure supporting the reuse of global identifiers for any entity named in the Web or in other network-based applications. LiveMemories might benefit from this infrastructure as a systematic way of ensuring that the information extracted about an entity is associated to the same ID, which is globally used on the Web at large.

8. Training activities for young researcher and technicians

As mentioned in Section 8, in recent years UniTN and FBK have established a number of joint training initiatives in the areas of HLTI and knowledge technology, from the Master level to the PhD, which provide a thorough grounding in the technology both from an industrial and from a research perspective. One of the strengths of the proposed project is that it will provide both applied and research-oriented training for a number of students and researchers by taking advantage of this curriculum. Whereas experienced researchers will take care of the delivery of the systems, the rest of the research will be carried out by PhDs in the DIT and CIMEC programs (for the longer-term projects), as projects in the Master, and by young technicians. (We expect between 20-30 PhDs, around 10-15 Master students, and perhaps as many young technicians.) Because the project involves both advanced research aspects and development aspects, and thanks to the involvement in the project of much of the local HLTI industry - which is already funding several fellowships in the HLTI Master - these young researchers and technicians will have the opportunity to get exposure not only to academic-style research but also to more applied work. Conversely, the participation of the local industry will give their personnel access to advanced developments.
We are also considering the possibility of a Summer School in HLTI to attract and evaluate potential students. This opportunity follows from the success of the EVALITA evaluation campaign run by FBK in 2007, where for the first time ever a considerable number of systems have been evaluated for five shared tasks (POS tagging, Word Sense Disambiguation, parsing, Named Entities recognition and temporal expressions recognition) for the Italian language. We think that having at alternating years an Evalita summer school, where students can practically experiment and prepare annotated data, and the Evalita workshop, with the evaluation campaign, would have a relevant impact on the state of art of human language technologies for Italian. Although the Evalita summer school would have a national/international organization, we have foreseen for the project a budget for helping the lunch of this initiative under the impulse of LiveMemories.

9. Potential impact on the social and/or economic context

LiveMemories intends to impact on the local context at several levels:

- A huge amount of memories of the local community will be collected, digitalized and stored, so as to guarantee their preservation. The network of local historical museums will be involved in all phases of this process.
- The Content Management Platform that will be delivered in Activity 4, is expected to be a significant step in the direction of a permanent service for monitoring facts about the territory from multiple information sources.

The target groups of the project involve almost everybody in Trentino, with an upper bound only in the ability that users will have in imagining and creating communities. Relevant target groups are:

- Trentino citizens who will be involved in a common effort to collect and share memories of the towns. The project will be a unique opportunity to stimulate communities to participate in a collective enterprise. We expect that sharing a common goal will overcome the perceived distance between institutions and street people.
- Tourists: the creation of tourist communities will make Trentino more interesting, with its unusual combination of high-tech and natural resources. This will bind tourists into a larger community and will increase their commitment to the territory.
- Students and schools, thus creating a future for our community and creating the premises for an active and proactive involvement of all citizens in a lifelong project. Schools will be actively involved in the project.
- Trentino companies, with the goal of exploiting the underlying multi-media material produced by the project, towards the development of specific vertical applications in areas such as tourism, e-government and security. The impact of the project will be magnified via a cultural debate, through the organization of scientific workshops and conferences and events open to the general public. We aim at bringing a cross-disciplinary debate which will be always organized at two levels: (i) selected topic among those developed by the project; (ii) the impact that a project like LiveMemories has on the topic itself, touching on issues like the role of Web-based technologies for Digital Memories, the sociological impact of communities, the new perspectives given by the Web Science Initiative. We plan to organize a public conference on these issues with Tim Berners Lee, inventor of the Web, and Wendy Hall, the only ERC commissioned with Computer Science expertise, both professors at SOTON.

Finally, we expect a positive impact on the research context of the territory from the collaboration among groups of FBK and UniTN. Particularly, in the HLT area, the groups at FBK, CIMEC and DIT are expected to establish strong research cooperation and share software components effectively.

10. Existing intellectual property rights and their impact on the results

In case the project proposal will be positively evaluated, before the start of the project, the coordinator (FBK) and the other partners (UNITN and SOTON) will agree on a contract, which will define the organization of the work among the parties, the management of the project, rights and obligations of the Parties, including their liabilities and property/access rights of the parties.

CORE TECHNOLOGIES AND RESOURCES
The IPRs for core technologies and resources used in the project are fully owned by the partners of the consortium. Most of the background knowledge (technologies already developed) and all foreground knowledge (developed during the project life) will be made available for research purposes in order to maximize diffusion and impact within the scientific community, as indicated below. The following is a list (still incomplete) of background knowledge available at the consortium:

- SWEB: a distributed infrastructure for the creation of document classifications, communities, and location and event clusters, and for semantic navigation ad search. All foreground knowledge developed on top of SWEB will be made publicly available.
- TextPro (basic text processing for Italian and English), MultiWordNet (lexical database for Italian) and I-CAB, a corpus of Italian annotated by FBK, are distributed for research purposes by FBK. The VENEX corpus of anaphoric reference in Italian, produced by UniTN, will be available to the partners.
- The ELKFED/BART toolkit for coreference is in the process of being made available in open source format through SourceForge.
- The Machine Translation technology (MOSES), is distributed as Open Source.
- Automatic Speech Transcription technology previously developed at FBK has been licensed to Pervoice, which will be involved as subcontracting for providing transcription services.

NON CORE TECHNOLOGIES
LiveMemories will need to integrate the core technologies with already existing platforms for managing geographical locations and communities. As for the first, a good candidate is Google Maps, for which public APIs are available, allowing easy integration in applications. As for community management, an option is to use the Flickr API, a platform developed by Yahoo and largely used; we will explore the possibility for a specific instantiation of Flickr for LiveMemories.

DATA FROM LOCAL PROVIDERS
The data used in the project (e.g. broadcast news from RAI) are subject to property rights imposed by the respective providers. Following a schema successfully applied during the Ontotext project, we intend to reach agreements with them for the availability of the data for research purposes during the life of the project. A potential source of risk is the availability of data from non institutional subjects (e.g. pictures form citizens). We intend

to take advantage of legal experts in the field since the very beginning of the project (related costs have been foreseen in the budget).

11. Project Consortium

11.1 FBK-irst

Description of scientific-technological competences

Fondazione Bruno Kessler (FBK), born on March 1st 2007, is a non-profit body with a public interest mission having private law status and inheriting the history of Istituto Trentino di Cultura (ITC – founded in 1962 by the Autonomous Province of Trento). Scientific excellence and innovation as well as technology transfer to companies and public services are FBK's main objectives. In its areas of competence, FBK collaborates with the main actors in global research and works in accordance with European Union Programs. The total budget is currently about 34 M Euro.

FBK - Scientific and Technological Research activities are carried out through three main areas: Information Technology, Microsystems, and Applied Physics. The research staff consists of about 80 people on a permanent basis, and about 50 people on temporary contracts. The budget amounts to about 19 M Euro. Half of the direct costs are covered by industrial contracts and European and National contracts. So far, FBK has carried on over 50 European contracts.

Within the Information Technology area, two Research Units are directly involved in the proposal, namely the Human Language Technology (HLT) Research Unit and the Data and Knowledge Management (DKM) Research Unit.

The first unit originated from three research lines of the former TCC and SSI Divisions of ITC-irst, while the second unit originated from a research line of the former SRA Division. The competence of the HLT unit is mainly in the following areas: automatic speech transcription, automatic speech recognition, extraction of linguistic information from audio data, cross-language information processing, automatic machine translation, speech interaction in noisy environment, information extraction (in particular, extraction of entities and relations between entities), question answering, word sense disambiguation, lexical acquisition and development of multilingual lexical resources. The HLT unit has developed state-of-the-art technology in all main research fields it contributes in and has performed consistently well in several international evaluations, such as DUC (summarization, best system for linguistic quality in 2005), SENSEVAL (word sense disambiguation, best unsupervised system for English all words, 2nd best supervised system for Spanish, Italian, Catalan), CLEF (cross language question answering, since 2004, best system for monolingual Italian/Italian), TREC (English question answering, since 2002, 4th position in 2003, first European system), PASCAL-RTE (recognizing textual entailment, two participations), TERN (temporal expressions, 2nd system on full TE recognition and normalization in 2004), NIST-MT (machine translation, since 2003, 4th on Arabic to English in 2006), IWLST (spoken language translation, best system in 2005).

The group working on text and speech translation is currently engaged in the most relevant open source project in the MT community. Research on speech recognition also meets the highest standards, and has reached the application market in several occasions.

Research on content extraction has a strong record of publications and evaluation results, namely on question answering and information extraction tasks.
Moreover, people of the unit are key-players of many international initiatives around evaluation and benchmarking (CLEF, PASCAL-RTE, IWLST, EVALITA). Finally, work on cognitive models focused on so called affective computing, both on text and speech sides, resulted in publications at top conferences.
The competence of the DKM unit concerns: knowledge representation and reasoning, semantic web technology for knowledge representation, knowledge acquisition, semantic matching of heterogeneous ontologies, logics for distributed reasoning, machine learning, semantic matching, and recommendation systems.

Description of the project's research group

The project group is composed of three main actors within FBK, namely the Human Language Technology (HLT) and Knowledge and Data Management (DKM) Research Units for the research activity and the Local Relationships group for the activity on the showcase. In the following, the project group is described focusing on the activities connected with the project.

HLT RESEARCH UNIT
Among all the research activities carried out within the HLT unit, we describe only those which are directly involved in the project.

Speech Recognition
This activity focuses on the following topics. (1) Core technology. The research includes: unsupervised cluster-based normalization of acoustic features; data selection methods for language modeling; detection and recognition of unknown words through sub-word units; word-graph expansion and re-scoring through morphology; and fast pronunciation modelling for new languages. (2) Speech Analytics. Work focuses on the extraction of informative and reliable prosodic features, on methods to evaluate compliance of spoken dialogues with service protocols, as well as methods to measure linguistic competences of scholars. (3) Multilingual Technology. So far, the multilingual technology developed covers three languages, namely Italian, English and Spanish.

Machine Translation
This activity covers research in the areas of text and speech translation. (1) Statistical Modelling. This research explores: (i) the integration of linguistic knowledge into the heart of the statistical translation engine; (ii) the use of context in document translation to improve coherence of translation; (iii) methods to cope with under-resourced translation pairs to overcome the data bottleneck of statistical methods. Some of the results are released as open-source (Moses, IrstLM). (2) Speech translation. Investigated issues are: (i) better interfaces between speech recognition and translation to reduce propagation; (ii) improving quality of speech translation output to better match human-made transcripts. (3) Knowledge Acquisition from Data.This topic includes: (i) adaptation of generic translation models to specific contexts/domains in order to focus on a specific topic; (ii) automatic acquisition of bilingual data from comparable; (iii) data-selection techniques from huge corpora to scale down complexity and fit specific application domains; (iv) methods to handle words in the source text that were not observed in the training data.

Content Processing
This activity covers research in the areas of question answering and content acquisition and integration from textual documents. The developed techniques are applied in concrete use cases provided by the users involved in the various projects. (1) Question Answering. Research activities focus on: (i) the development of an entailment-based approach to closed domain QA, and (ii) the participation in international open domain QA evaluation campaigns (both monolingual and cross-lingual tasks).
Application oriented activities aim at delivering publicly accessible QA demonstrators (with mobile and desktop interfaces) based on a distributed Web Services architecture. (2) Knowledge Acquisition. The issues investigated within this research area include: key-phrase extraction, extraction of both relevant entities and relations between entities, ontology learning and population, integration of textual techniques within a framework for multi- and cross-media knowledge acquisition, automatic acquisition of textual entailment patterns. In this task both rule-based and machine-learning techniques are applied. (3) Content Integration. The activity focuses on: (i) intra document and cross document co-reference of entity names (persons, organizations and locations), (ii) co-reference of relations across time. The output of the integration process is used to design and implement innovative information access systems according to Semantic Web guidelines.

Infrastructure
This activity provides technological support and high-level services in order to optimize the activities of the HLT Research Unit. (1) Technological infrastructure. It includes the management of the cluster of high performance machines, the installation and management of specific (e.g. linguistic) software tools and packages, the storage and retrieval of huge data (e.g. acoustic data) as well as the definition of their format and documentation, the support for inter-process communication of research prototypes, as well as the infrastructure to support Web-based demonstration systems. (2) Language Resources. The linguistic infrastructure is in charge of two main activities, namely (i) development and maintenance of written, spoken, and multilingual resources, and (ii) networking and dissemination. The first includes design and data collection, definition of annotation schemes, automatic annotation, creation of training data and gold standards, maintenance and management. The networking activity mainly focuses on maintaining relationships with other institutions, distributing resources, disseminating results, organizing events and evaluation campaigns.

DKM RESEARCH UNIT
The research activity consists of developing logical and computationally sustainable models to support knowledge creation from data, knowledge integration, and reasoning services. (1) Knowledge elicitation.
We concentrate on the process of knowledge extraction from data, such as preprocessed textual data (e.g., tagged textual and multimedia documents), data with a light semantics (e.g. folksonomies and wikies), structured data (e.g., database and XML files). The result of knowledge elicitation is a machine processable logical theory, called knowledge module. (2) Knowledge integration. We concentrate on three main factors: heterogeneity (i.e. data stored in different knowledge modules can have different semantics), autonomy (i.e. the management of knowledge modules cannot be centralized), and scalability (i.e. the dimension of single modules and the number of modules can be extremely large). (3) We develop logical and heuristic reasoning services, and we apply them to support: automatic content extraction and integration, semantic web service composition, and analysis of medical procedures.

LOCAL RELATIONSHIP GROUP

The Local Relationships group is in charge of maintaining all the relationships with the numerous FBK stakeholders. The 40 years long various and consolidated FBK activity, ranging from scientific and technological research to the humanities, has lead to the creation of a vast network of relationships. The goal of the Local Relationships group is to make FBK relationships with other research institutes, as well as public institutions and private companies, concrete and effective. Fund raising, stakeholders' needs and expectations monitoring, dissemination of research results, enhancement of FBK consultancy activity on the territory, are among the objectives of the Local Relationships group.

As regards specifically the research projects, the Local Relationships group aims at actively involving all those actors which can benefit from the collaboration with FBK on different dimensions, such as the growth of the culture of innovation, the most strategic aspects of the international scientific research, the identification of the development potential of technological applications.

Connection with own research programs and positioning regarding own research strategy

The advanced research carried out and the technology produced until now will be exploited by the FBK group to cope with the new research challenges posed by the project.

According to their own research programs, the three components of the FBK group will focus on different activities. The HLT unit will be responsible for the activity on content extraction from multimedia sources, will work on coreference and geo-time tagging (content integration activity) and will be in charge of the creation of the LiveMemories platform. The DKM unit will be responsible for the activity on content integration and the Local Relationships group will be in charge of the showcase.

As regards the research strategies, through the project FBK will open new research activities in the Human Language Technologies and Semantic Web fields. More specifically, a new research line will be activated in data driven content integration, where statistical techniques already successfully applied to tasks in HLT (e.g. named entity recognition) will be experimented on a number of higher level phenomena.

FBK will also maintain the consolidated network of strategic collaborations involving key players in the territory; among them the University of Trento (Master school in HLTI, Doctorate school in ICT), the University of Bolzano (Master in LCT); UniTN-CIMEC (joint research on language and cognition, planning of degree in Cognitive Informatics), CELCT (organization of evaluation campaigns) and PerVoice (spin-off of HLT's speech recognition technology).

Besides contributing from the point of view of research, FBK will also have the coordination of the Content Integration Platform, a joint effort with UniTN to build a multimodal digital library for content acquisition and integration. This activity will exploit and enhance the competence of the HLT unit infrastructure service.

11.2 University of Trento

Description of scientific-technological competences

The University of Trento has occupied the top rankings in the national ranking of Italian Universities and Faculties and the first place in many scientific and engineering areas since 2001, when the national ranking by CENSIS began.

DIT

The Department of Information and Communication Technology (DIT), although only created in 2002, has an outstanding scientific record and an outstanding performance at attracting R&D funding from both industry, local government, and the EU. In the 6th Framework, DIT participated in 22 projects even though it only became independent in the second part of the 6th Framework. Between 2004 and 2007 DIT acquired over 16M Euro of research, industrial and educational funding, while support form the university in this period was 6M Euro. DIT has also developed highly effective training curricula which have quickly achieved a high visibility. The International PhD program in ICT has been an enormous success; currently, there are 155 PhD students in the program. DIT currently has 74 members of academic staff, 24 members of research staff, 14 PostDocs, 10 members of technical staff, and 19 members of administrative staff. DIT is historically strong in the area of knowledge representation and reasoning and has been quickly expanding in the areas of Web Science, Human Language Technology, and Interfaces.

CIMEC

The Center for Brain / Mind Sciences (CIMEC) was recently set up to put Trentino at the forefront of interdisciplinary research at the frontiers of Neural Science, Psychology, Computer Science, and Physics, and is quickly establishing itself as one of the leading labs of this type in Europe; its International PhD in Cognitive and Brain Sciences is also already attracting a great degree of high-quality students from Italy and abroad. The Language, Interaction and Computation Lab is one of the four labs around which the activity of CIMEC is articulated. It consists of 5 faculty, 2 postdocs, and 7 PhD students, and its research focuses primarily on the acquisition of commonsense knowledge from very large amounts of data and on adaptive interfaces.

DSSR

The Department of Sociology and Social Research is committed to the investigation of macro-areas such as social structure; inequalities and collective actions; social norms, political and ethical values; social policies and European institutions; culture and organizational change; and social theory, and is deeply engaged in research projects and activities at a national and international level. Based at the Department, the Research Unit on Communication, Organizational Learning and Aesthetics (www.unitn/rucola) consists of scholars and researchers collaborating since 1993 on the basis of common professional interest in aspects of Organization Studies and Information Systems and a bias in favour of qualitative and interpretive research methods. The unit includes full time faculty members, independent researchers, and doctoral students.

Description of the project's research group

The groups from UniTN that are involved in the proposal are internationally known for their work in all areas of Web Science, Human language technology and Interfaces, and Knowledge Management and reasoning that will be central to the project.

Web Science is a new but rapidly emerging area of growth for UNITN. In the Web Science area DIT has engaged in
- The development of efficient multi-tier web-based systems, both traditional and Web 2.0/Semantic Web, and including Web Service technology (e.g. project ELEAF).
- The development of languages for knowledge representation in the Web. One contribution, joint with ITC-Irst has been the development of C-OWL, a language for the interconnection of multiple ontologies (written in OWL).

DIT is also the coordinator of the LiquidPub initiative (http://project.liquidpub.org/) whose goal is to bridge and formalize the social and technical aspects of the paradigm of publishing of scientific material on large-scale networks like the web. The current focus of the initiative is the development and promotion of a new paradigm for the way scientific knowledge is produced, disseminated, evaluated, and consumed. The paradigm, powered by the advent of the Web and advances in ICT, introduces the notion of Liquid Publications, which are evolutionary, collaborative, and composable scientific contributions. Many Liquid Publication concepts are based on a parallel between scientific knowledge artefacts and software artefacts, and hence on lessons learned in (agile, collaborative, open source) software development, as well as on lessons learned from Web 2.0 in terms of collaborative evaluation of knowledge artifacts. For more information, see project site at http://project.liquidpub.org/.

Human Language Technology and Interfaces has been in recent years one of the main areas of growth for UniTN. A key element of UniTN's strategy in this area has been exploiting the synergy with IRST, which this proposal will greatly strengthen. UniTN's first objective has been to put in place a curriculum to train individuals in these areas—both those oriented towards an academic career, and those with an inclination towards industry—and to attract students from abroad. In addition to the existing and very successful International PhD program in ICT at DIT, with a technological focus and with a strong presence of IRST, in the last two years UniTN has thus first developed an International PhD program centered at the Center for Brain / Mind Sciences in Rovereto with a focus on the cognitive aspects, a strong language and interaction component and that will also emphasize the interface aspect. A new Master of HLT and Interfaces is starting the current academic year, a joint initiative with IRST, conceived both to train personnel for the continuously growing local industry in the area and to prepare those students with a more academic bent for the PhDs. These initiatives have been supported through the hiring, so far, of five faculty and have encountered a warm response in terms of student applications and support from the local industry, particularly Cogito and PerVoice, that are supporting the courses with fellowships and the collaboration with whom we also expect to strengthen through this project. New initiatives are in preparation, which will further strengthen the area.

In the HLT area, the groups from DIT and CIMEC have established a reputation of excellence for their work in text processing, speech processing, statistical machine translation, and interfaces, including:

- The creation and usage of natural language processing tools for tokenization, part-of-speech tagging, lemmatization;
- Text mining, including work on named entity extraction, relation extraction, ontology learning, and semantic role labeling, also through participation in international competitions (CoNLL, etc);
- Entity disambiguation and its applications in HLT (members of the UniTN team organized the highly successful 2007 Johns Hopkins workshop on coreference and entity disambiguation);

- The creation, pre-processing and linguistic annotation of corpora ranging from the high-quality hand annotation of medium size corpora (GNOME, ARRAU, LUNA) to the automatic annotation of very large corpora (e.g., as part of the WaCky corpora international initiative);
- The development of machine learning methods and their practical application in HLT, also through the development of tools (e.g., active learning, kernel methods);
- Very large vocabulary speech recognition research work evaluated in international evaluation tests (e.g. DARPA ATIS and EARS). Such systems were in the top three amongst US and European research labs;
- Spoken Language Understanding research and interfaces to databases (e.g. ATIS). Such system scored first in the international DARPA evaluation tests amongst US and European research labs. UNITN is involved in the EC-funded LUNA project on multilingual spoken language understanding;
- Statistical machine translation for very large text bilingual corpora (e.g. French-English and Arabic-English) based on finite state machine tools (DARPA TIDES project);
- Human-machine interfaces ranging from multilingual conversational systems for mixed initiative dialog supporting spontaneous speech input (e.g. "How May I Help You?" project) to web-based interfaces. The UNITN team was awarded a Marie Curie Research Excellence grant to work on next generation human-machine interfaces.

Knowledge Management, the knowledge and reasoning group at DIT is internationally known for their work on:

- Contextual reasoning and the maintenance of multiple perspectives on knowledge;
- Lightweight ontologies, providing a theory of how to translate standard classifications, such as DMoz, into formal classifications, namely graph structures where labels are written in a propositional concept language. This allows reducing essential tasks on classifications, such as document classification and query answering, to reasoning about subsumption;
- Semantic Matching, viewed as an operation which takes two graph-like structures (e.g., web classifications, business catalogs, ontologies) and produces a mapping between the nodes of these graphs that correspond semantically to each other. Our approach allows for a translation of the matching problem into a propositional validity problem, which can then be efficiently resolved using (sound and complete) state of the art propositional satisfiability solvers.

As well as their work on efficient automated reasoning using modal and temporal logics, fast SAT based decision procedures; peer-2-peer data and knowledge management; entity centric knowledge management; and logics for knowledge representation.
Communication, Organizational Learning and Aesthetics RUCOLA is internationally known for their work on the social construction of technology; the exploration the practices of organizing; a focus on knowing and learning as a collective, social, affective and not entirely cognitive activity; and a strong emphasis on the relation between linguistic, symbolic, material and emotional aspects of organizational processes.

Connection with own research programs and positioning regarding own research strategy

UniTN is motivated by the opportunities offered by the move to a 'communal' not exploited by the methods for sharing and labeling knowledge embodied in Flickr, Google Earth, etc. Documents shared in this way—both language and images—contain a wealth of information that is currently not exploited even for retrieval, let alone to give users new perspectives on this information, or to extract the commonsense knowledge that is so crucial for many HLT applications.

In the area of knowledge management, research is moving away from global schemes for organizing knowledge defined in advance, and towards user-defined classification schemes for knowledge representation and data management. This requires the developing methods for: (i)ontology building from user classifications through NLP; (ii)reasoning on user ontologies to support and automate tasks such as content classification, semantic search, detection of inconsistencies; (iii)ontology matching to find correspondences between ontologies of different users; (iv)analyzing and exploiting such correspondences.

Our current research in HLT focuses (i)processing and extracting knowledge – in particular, relation-based entity descriptions – from very large amounts of data; (ii)ways of organizing this knowledge, e.g., via entity disambiguation; and (iii)moving from text and speech knowledge to integration with knowledge extracted from images and videos (the interaction with Southampton will be of great benefit in this).

In the area of Interfaces, this project is tied to our work on (i)ways of organizing and accessing knowledge (e.g. through semantic search) and of displaying it in an adaptive way (e.g., through narratives)–again, taking advantage of Southampton's experience; and (ii)ways of ontology visualization and management (e.g., populating or changing an ontology). The challenge here is how to make it a natural task reducing the learning curve of ontologies.

11.3 University of Southampton

Description of scientific-technological competences

School of Electronics & Computer Science (ECS)

ECS is the UK's leading School in the field, with the top 5* rating for research in the last two Research Assessment Exercises, and awarded the 'best 5* rating by the Higher Education Funding Council for England in 2003. It has a strong presence in many relevant areas; it organized the 2006 World Wide Web conference, was the first academic institution in the world to adopt a self-archiving mandate, and created the first and most widely used archiving software EPrints. Former Head of School was Professor Wendy Hall. The Intelligence, Agents, Multimedia group is an interdisciplinary group focusing on the design and application of computing systems for complex information and knowledge processing tasks, with expertise in areas including grid computing, peer-to-peer systems, sensor networks, the Semantic Web, and pervasive computing environments. Tim Berners-Lee, director of the World Wide Web Consortium is a Professor in the group. ECS took in £22m in research income in 2004-5, published 642 papers in that year.

Web Science

The Web Science Research Initiative (WSRI) is a joint endeavour between MIT's CSAIL and ECS, with four directors: Tim Berners-Lee, Wendy Hall, Nigel Shadbolt, professor of AI at Southampton and Director of AKT (2000-2007) and Daniel Weitzner, Technology

and Society Domain leader of W3C and principal research scientist at MIT. James Hendler, Professor of computer science, is Associate Director.
WSRI is pioneering Web Science, to study decentralized information systems. Web Science involves engineering new infrastructure protocols and understanding the society that uses them, to support reuse of information in new contexts. It focuses on decentralization to avoid social and technical bottlenecks, and integrates powerful scientific and mathematical techniques from many disciplines to consider at once microscopic properties, macroscopic phenomena, and their relationships. IAM's research in this area has focused on the ability to keep information and links separately (open hypermedia) so that it can be personalized to each user (adaptive hypermedia) using sophisticated models of the meaning of documents, data and their interconnections (the Semantic Web). The group is in the forefront of Semantic Web development to shift the focus of the Web from documents to data.

Memories for Life
Memories for Life (M4L) is a network, funded by the EPSRC (2004-7) to bring together a range of academics in a bid to understand how memory works and to develop the technologies to enhance it now that human memory is supplemented by increasing amounts of personal digital information; emails, photographs, Internet telephone calls, GPS locations and television viewing logs. M4L brought together psychologists, neuroscientists, sociologists and computer scientists to investigate effective use and management of both human and computerized memory.

Multimedia Imaging
IAM has a strong reputation in multimedia research, such as content analysis of a range of multimedia including images, video, audio and augmented reality, to allow more powerful content-based browsing, retrieval and navigation, and give direct access to the semantics. It is also investigating multimedia systems architectures to combine the automatic extraction, representation and manipulation of semantic content with more traditional information handling facilities in distributed environments. Another focus is on tools and techniques to facilitate from video databases and the interpretation and presentation of that information with hypermedia. Current work includes high resolution image retrieval and navigation systems for art gallery collections, the handling of continuous metadata for continuous media streams, the development of content-based retrieval of 2D/3D and the combination of images and 3D scans to characterize art works.

Description of the project's research group

IAM is one of ten research groups in Electronics and Computer Science at Southampton. IAM's research focuses on complex information and knowledge processing tasks, with various themes including agents, grids and knowledge technologies. The group investigating LiveMemories has expertise in three particular areas: knowledge technologies, multimedia and Web Science (the science of decentralised information systems).

Knowledge Technologies
Knowledge technology is a major theme in IAM, focusing on the technologies, formalisms and protocols of the Semantic Web. In a major project, Advanced Knowledge Technologies (AKT), IAM was a lead partner.

AKT conceptualised the knowledge lifecycle as a series of six stages, ranging from knowledge acquisition, modelling, retrieval, reuse, publishing and maintenance. The group developed and extended technologies and standards to provide integrated methods and services for these tasks.
The AKT approach uses Semantic Web technologies to provide a means of interrogating resources for content; intelligent systems support the SW, and the SW primes, builds and drives intelligent systems. This research led to several intelligent tools and services for creating and maintaining content on the Web, and for collaborating in the creation of documents, datasets and other knowledge-based resources.
Another important line of work was that of creating the marked-up content in the form that the SW requires, using methods ranging from Natural Language Processing (NLP), ontology-mediated knowledge harvesting, detecting duplicates and ensuring referential integrity, automatic and semi-automatic markup and tools for aiding annotation (including multimedia). Although Southampton's research was not in the NLP field, it collaborated closely with NLP researchers from other universities, and is extremely receptive to the importance of the use of NLP techniques on the Web, and to manage large knowledge repositories.
Tools were developed in collaboration to mediate analysis of texts using ontologies to produce filled templates and services such as annotation for information extraction. The scale of the content available on the Web has allowed us to go beyond user-centred annotation, towards unsupervised and semi-supervised extraction.
AKT also developed important pieces of infrastructure both for the Web and for large-scale information repositories, at all levels, from scalable data warehousing technologies to interfaces, including methods and tools for searching and browsing, and tools to find content and also to make sense of the content retrieved. The methods developed have and will also be applied to other large knowledge stores, including multimedia and organisational archives, facilitating the reuse of knowledge bases. Furthermore, this infrastructure has been exploited in a spinoff firm dedicated to ensuring the security of data and privacy of its owners – a key issue in a public repository of memories.
One particular strength is the development of tools, methods and techniques for working with ontologies – often regarded as a potential bottleneck for the development of the SW. AKT's work on ontologies includes methods for search, evaluation, mapping, merging, maintenance, modularisation, fragmentation and pruning.
Another important strand of AKT research involves using SW information to induce human, institutional and organisational relationships. AKT tools have been developed to identify communities of practice of individuals, to locate expertise and to associate individuals with their expertise. The IAM group also contains a number of researchers who have investigated issues of trust and provenance of information, issues that come to the fore in the context of large distributed repositories.

Multimedia
Systems for storing, retrieving, browsing and processing digital multimedia information are increasing in number rapidly as facilities for capturing digital images, digital audio and digital video become cheaper and easier to use. The development of versatile facilities for effective handling of this material offers many unsolved research opportunities and challenges. Real-time processing enables direct support of people interacting with their environment, so that images, video and audio can be used to link digital and physical worlds.
IAM's research involves content analysis of a wide range of media including images, video, audio and augmented reality. It aims to build systems to provide more powerful

content based browsing, retrieval and navigation. More intelligent facilities are being investigated to give direct access to the semantics of the media. This involves the solution of many problems in computer vision, audio (speech and music) understanding and 3-D visualisation. Examples of current work within the group include the development of high resolution image retrieval and navigation systems for art gallery collections, the handling of continuous metadata for continuous media streams, the development of content based retrieval of 2D/3D and the combination of images and 3D scans to characterise art works.
Combining the automatic extraction, representation and manipulation of semantic content with more traditional information handling facilities in distributed environments requires new multimedia system architectures and these are also being investigated. The development of integrated cross-media content and concept based browsing, retrieval and navigation for distributed multimedia is continuing to provide a challenging long-range research goal.
Knowledge technologies and multimedia were combined in the AKT project with the development of Photocopain, a system designed to address the resource-intensive, increasingly essential task of describing and archiving personal experiences, where casual users are unwilling to expend large amounts of effort on creating the annotations which are required to organise their collections. Photocopain is a semi-automatic image annotation system which combines information about the context in which a photograph was captured with information from other readily available sources in order to generate outline annotations for that photograph that the user may further extend or amend.

Web Science
The Web is one example of an increasingly common type of information structure, the decentralised and distributed information store that presents users with a means of accessing an ever increasing number of diverse information sources (ranging from institutional archives to real world deployed sensor networks).
However, using of this information to make informed decisions presents a number of challenges. The information may be incomplete, uncertain or contradictory. It may come from sources owned by different stakeholders, and increasingly, it may incorporate a huge number of different media. Thus, there is a clear need for systems that are able to interact with these diverse information sources to not only collect relevant information, but also to reformulate and reason about it in principled ways.
The Web is defined by a few simple rules which give rise to highly complex structures. Web Science is intended to get to grips with the relationships between the simple rules and the complex behaviour, to identify trends that could threaten or fragment the Web, and to contribute to the research required to ensure its continuing development. IAM's research has focused on the ability to keep information and links separately (open hypermedia) so that it can be personalised to each user (adaptive hypermedia) using sophisticated models of the meaning of documents, data and their interconnections (the Semantic Web).
Many systems that the Group has built have been used for engineering, publishing and education, and have helped shaped current Web standards. Meanwhile, the group has always been in the forefront of development of the Semantic Web as a means of shifting the focus of the Web from documents to data.

Connection with own research programs and positioning regarding own research strategy

The research in the LiveMemories project complements several IAM research themes, including those of knowledge technologies, content extraction using Semantic Web-style technologies, and memories for life.

By providing an example of a large decentralised information system, with content provided in a distributed fashion, the project also acts as a testbed for IAM's Web Science activities.

The use of knowledge technologies in a multimedia environment (Photocopain, as described above) was a very successful marriage of technologies, but this particular system was a small part of the AKT collaboration, and so was a proof of concept. The integration between knowledge technologies and multimedia should be explored in more detail in this forum, to understand further the possibilities of other available information (such as ontologies, folksonomies, tags and annotation of related resources, and content extracted from related resources). Furthermore, available information could be used in a more integrated way. In this respect, the project would help create synergies within the group between multimedia research and knowledge technologies, as well as those studying the dynamics of large-scale decentralised information repositories. The area will also act as a testbed for more straightforward image analysis techniques being used at IAM, including machine learning and various type of feature detector, which would be used as subtools.

The Memories for Life subfield, in which IAM hosted an EPSRC network, is an area where the practicality of IAM research will be highlighted. As information storage becomes ever-cheaper, navigating very large repositories will become pressing, not just in the business context, but in ordinary life, and this is a research area in which IAM wishes to maintain its current prominent position.